PEER REVIEW, Vol. 6, Nº 13, 2024
DOI: 10.53660/PRW-2346-4321
ISSN: 1541-1389

Percurso metodológico para implementação do linkage nos sistemas de informação em saúde brasileiros

Methodological process for implementing linkage in Brazilian health information systems

Pauliana Valéria Machado Galvão
ORCID: https://orcid.org/0000-0002-4418-218X
Faculdade de Ciências Médicas, Universidade de Pernambuco, Brasil
E-mail: pauliana.galvao@upe.br
Paloma Luna Maranhão Conrado
ORCID: https://orcid.org/0000-0001-8828-667X
Faculdade de Medicina, Campus Serra Talhada, Universidade de Pernambuco, Brasil
E-mail: paloma.luna@upe.br
Patricia de Moraes Soares Santana
ORCID: https://orcid.org/0000-0001-7448-0623
Faculdade de Medicina, Campus Serra Talhada, Universidade de Pernambuco, Brasil
E-mail: patricia.santana@upe.br
Fernando Antônio Ribeiro de Gusmão Filho
ORCID: https://orcid.org/0000-0002-9255-0760
Faculdade de Ciências Médicas, Universidade de Pernambuco, Brasil
E-mail: fernando.gusmao@upe.br
Polyana Felipe Ferreira da Costa
ORCID: https://orcid.org/0000-0002-6054-8401
Faculdade de Medicina, Campus Serra Talhada, Universidade de Pernambuco, Brasil
E-mail: polyana.costa@upe.br
Carolina Maria da Silva
ORCID: https://orcid.org/0000-0003-2680-9767
Faculdade de Medicina, Campus Serra Talhada, Universidade de Pernambuco, Brasil
E-mail: carolina.silva@upe.br
George Alessandro Maranhão Conrado
ORCID: https://orcid.org/0000-0001-6649-577X
Faculdade de Medicina, Campus Serra Talhada, Universidade de Pernambuco, Brasil
E-mail: george.maranhao@upe.br

RESUMO

Os Sistemas de Informação de Saúde são componentes empregados na coleta, processamento, análise e distribuição da informação de saúde, registrado de forma contínua pelo Ministério da Saúde para fins administrativos e de vigilância. Quanto mais integrados e interoperáveis, maior a eficiência deste sistema. Assim, este artigo pretende relatar o percurso metodológico para implementação do linkage nos sistemas de informação em saúde, considerando as particularidades de cada sistema epidemiológico e empregando um software gratuito e o banco de dados anonimizado que é disponibilizado pelos canais oficiais.

Palavras-chave: Sistemas de Informação; Vigilância; Linkage

Recebido: 10/06/2024 | Aceito: 12/07/2024 | Publicado: 15/07/2024

ABSTRACT

Health Information Systems are components used to collect, process, analyze and distribute health information, recorded continuously by the Ministry of Health for administrative and surveillance purposes. The more integrated and interoperable these systems are, the greater their efficiency. Therefore, this article aims to report on the methodological path for implementing linkage in health information systems, considering the particularities of each epidemiological system and using free software and the anonymized database made available through official channels.

Keywords: Information Systems; Surveillance; Linkage

INTRODUÇÃO

No Brasil, dados secundários são amplamente disponíveis para a Epidemiologia (Areco *et al.,* 2021). Geralmente, eles são sistematizados nos Sistemas de Informação em Saúde (SIS), que são instrumentos de suporte à produção de informações em saúde, através do processamento de dados coletados em serviços públicos e outros locais (Coelho Neto e Chioro, 2021). A vigilância da saúde pública baseada na análise de dados desempenha um papel crucial na detecção e resposta a crises de saúde pública (Zhang *et al.*, 2024).

No entanto, os SIS foram construídos de forma independente com finalidades próprias e sem a obrigatoriedade da interoperabilidade (Areco *et al.*, 2021; Drumond *et al.*, 2009). Além disso, apesar das estratégias de ciência aberta e transparência de dados que faz o Ministério da Saúde brasileiro disponibilizar versões anonimizadas de dados individualizados para amplo uso em pesquisas, as interfaces utilizadas (Tabwin e Tabnet) não são tão "amigáveis" e apresentam limitações consideráveis (Petruzalek, 2016; Saldanha, Bastos e Barcellos, 2019). Basicamente, o Tabwin serve para tabular e tratar dados e o Tabnet é uma extensão do mesmo para promover tabulações mais rápida com dados resgatados diretamente da Internet (Silva, 2009).

As barreiras para o uso dos dados secundários vão desde a limitação dos dados a determinadas áreas geográficas ou períodos, mudanças na forma de coleta das variáveis, falta de padronização no formato dos dados, descontinuidade na coleta de alguns dados ao longo do tempo ou variação na cobertura (Areco *et al.*, 2021). Mesmo com tudo isso, a oportunidade do uso destes dados não pode ser desperdiçada. Assim, este estudo propôs a relatar um percurso metodológico para integrar os principais SIS com caráter epidemiológico através de um programa estatístico disponibilizado livremente.

A IDEIA

A partir da demanda tracada pela Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco para pensar a Saúde da Mulher (Edital n. 07/2022), o projeto entitulado **Linkage probabilístico dos sistemas de informação de saúde para vigilância da saúde da mulher pernambucana** foi proposto.

A proposta deste projeto foi de integrar os SIS e estabelecer um processo de vigilância da Saúde da Mulher pernambucana, com o uso de dados públicos. Um dos objetivos iniciais foi relatar o percurso metodológico utilizado neste processo.

PERCURSO METODOLÓGICO

Inicialmente, os bancos de dados dos Sistemas de Informações de Saúde foram adquiridos da página do DATASUS em seu formato anonimizado. O Sistema de Informação sobre Nascidos Vivos (SINASC) foi mantido integralmente para relacionar as mortes maternas com os nascimentos e mortes fetais. Os demais sistemas (Sistema de Informação sobre Mortalidade – SIM e Sistema de Informação de Agravos de Notificação – SINAN) foram trabalhados apenas para a população feminina. Foram traçadas estratégias específicas quanto a cada tipo de banco de dados empregado que serão demonstradas nas seções seguintes.

O programa estatístico utilizado foi o R, empregando os pacotes tidyverse e reclin.

Em linhas gerais, a primeira etapa foi realizada com a limpeza e organização dos sistemas de informações utilizadas. As variáveis de ligação variaram de acordo com cada banco utilizado.

Ademais, se fez necessário padronizar campos que foram utilizados para a vinculação. Por exemplo: A idade é trabalhada com 2 subcampos, o primeiro indicando a unidade da idade (0 = idade ignorada; 1 = horas; 2 = dias; 3 = meses; e 4 = anos) e o segundo subcampo a depender do sistema: O SIM e o SINASC indica a quantidade de unidades composto por 2 dígitos; o SINAN indicando a unidade da idade com 3 dígitos.

Foi necessário também identificar uma variável a ser usada como blocagem para minimizar o custo com o processamento dos dados, reduzindo a quantidade de pares possíveis e a perda de pares verdadeiros.

A segunda etapa para fazer o linkage é a “deduplicação” das bases de dados. Esse procedimento consiste em identificar e remover as duplicidades dos dados. Optou-se por fazer este procedimento de forma automatizada, empregando os códigos do linkage na comparação dos mesmos bancos em busca de repetições.

Na sequência, procedeu-se o linkage determinístico, onde todas as variáveis retornam uma correspondência exata entre os dados, e o linkage probabilístico, onde as variáveis elencadas são empregadas para calcular as similaridades de cada para apartir de um ponto de corte estabelecido previamente pelo pesquisador, optando-se pela distância de Jaro Winkler, no nível de 0,9.

Por fim, a revisão manual dos pares duvidosos visando a classificação dos mesmos como pares verdadeiros ou não pares foi realizada.

LINKAGE DO SIM E SINASC

Usando como exemplo um estudo onde associamos as mortes maternas com os nascidos vivos, esses bancos foram relacionados. Devido a questões de memória computacional, a estratégia mais eficiente foi relacionar cada banco anual.

Para o ano de 2021, 126.130 nascimentos e 77 mortes maternas para o Estado de Pernambuco, resultando em uma razão de mortalidade materna de 61,0 mortes por 100 mil nascidos vivos.

Um passo que antecede a vinculação dos dados é a definição das variáveis a serem utilizadas para esta técnica (Garcia, Miranda e Sousa, 2022). Considerando as variáveis referentes à mãe presentes no SINASC e as variáveis presentes no SIM, temos a **idade**, **estado civil**, **escolaridade**, **ocupação**, **raça/cor**, **código de município de residência** e **município de nascimento da mãe**. Assim, os nomes das variáveis foram adaptados para coincidirem e inicialmente foi realizado o linkage determinístico. Outro cuidado foi a criação de uma chave para permitir uma manipulação mais segura do dado.

A rotina determinística baseou-se na comparação dos registros (linkage interno à base de dados), empregando-se chave determinística composta pelas informações maternas e do nascimento, seguido da comparação automática dos códigos de município de residência e naturalidade, seguido de revisão manual (Aguiar *et al*., 2020).

Para este passo, a comparação de um banco com ele mesmo (deduplicação) foi realizada para remover dados repetidos, não sendo encontrada repetições.

A segunda comparação já considerou o banco de dados do SIM e SINASC. A Intenção é encontrar pares de combinação nos dois bancos, e destes devem selecionar os pares únicos. Podem ser feitas diversas rodadas para garantir a captação de pares possíveis. Na primeira rodada, 33 pares foram encontrados, 9 sendo pares únicos (11,7% das mortes maternas). O código empregado para realizar o linkage determinísticofoi apresentado no Quadro 1.

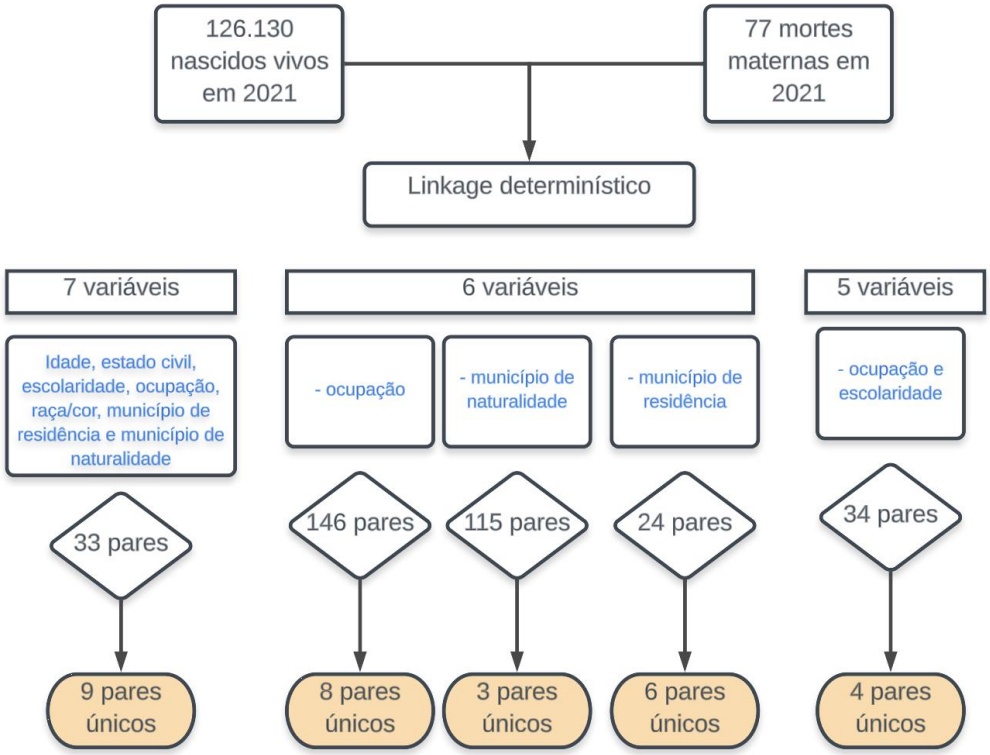
Quadro 1 – Códigos empregados para realizar o linkage determinístico do SIM e SINASC

```
#Criar um registro único nas mortes maternas
mm$chavemae<-seq(1, 77, by = 1)

#Criar um registro unico das notificacoes de sifilis materna
nv2021$chavenasc<-seq(1, 126130, by = 1)

#Primeiro linkage
pares_blocagem <- pair_blocking(x = mm, y = nv2021, blocking_var = c("codmunres")) |>
  filter_pairs_for_deduplication() #filtrando os pares duplicados
  pares_blocagem #visualizando a tabela resultante

p_deter <- pares_blocagem|>
  compare_pairs(by = c("id", "eciv", "esc", "ocup","raca", "codmunres", "codmunnatu"))

pares_iguais <- p_deter |>
  as_tibble() |>
  filter(id & eciv & esc & ocup & raca & codmunres & codmunnatu)

a<-table(pares_iguais$x)
a<-data.frame(a) #para não precisar ficar olhando no excel quantos tem cada um
b<-a[a$Freq == 1,]
b$Var1<-as.numeric(as.character(b$Var1))

registros_remover<-b$Var1
pares_iguais1<-pares_iguais %>%
  filter((x %in% registros_remover))

dados1<-mm %>%
  filter(!(chavemae %in% pares_iguais1$x))

link1mae<-mm %>%
  filter((chavemae %in% pares_iguais1$x))

#remover dos nascidos vivos
nasc1<-nv2021 %>%
  filter(!(chavenasc %in% pares_iguais1$y))

link1nasc<-nv2021 %>%
  filter((chavenasc %in% pares_iguais1$y))

link1mae$chavenasc<-link1nasc$chavenasc
link1<-inner_join(link1nasc, link1mae, by = "chavenasc")
```

Fonte: Pesquisa própria (2024)

Este passo foi repetido com a retirada de variáveis, para garantir a exploração de todas as possibilidades. O fluxograma abaixo ilustra a estratégia seguida:

Fluxograma 1: Linkage determinístico aplicado à associação de SINASC e SIM.



Fonte: Pesquisa própria (2024)

Após determinar os 30 registros vinculados a partir do linkage determinístico, os registros restantes foram vinculados pelo linkage probabilístico.

Da mesma forma que o linkage determinístico, a blocagem é feita com os registros restantes. Após esta etapa, a função compare_pairs() foi usada para especificara as variáveis a serem comparadas, utilizando a distância de Jaro Winkler para calcular a similaridade em cada uma das variáveis. O ponto de corte é de escolha do pesquisador, e por tal motivo optamos por usar 0,95.

O resultado gerou escores por variáveis e par a decisão de pares possíveis foi calculado um escore único que foi comparado a um ponto de corte calculado a partir do número de varíaveis empregados (no caso, ampliou-se as variáveis para nove, com a inclusão das variáveis data e mês, ficando 0,95 x 9 = 8,5). O script empregado foi apresentado no Quadro 2.

Quadro 2 – Códigos empregados para realizar o linkage probabilístico do SIM e SINASC

```
pares_blocagem <- pair_blocking(x = dados5, y = nasc5, blocking_var = c("codmunres")) |>
  filter_pairs_for_deduplication()
pares_blocagem

pares_link_prob <- pares_blocagem |>
  compare_pairs(by = c("id", "eciv", "raca", "codmunnatu", "codmunres", "esc", "mes", "ocup",
"dt"),  default_comparator = jaro_winkler(threshold = 0.95))
pares_link_prob

p3 <- pares_link_prob |>
  score_simsum() |>
  select_threshold(threshold = 8.5)

p3

p3 |>
  as_tibble() |>
  count(select)
    <lgl>   <int>
1 FALSE  366327
2 TRUE       17
```

Fonte: Pesquisa própria (2024)

Este Código gerou mais 17 pares possíveis diferentes dos 30 já encontrados anteriormente. O poder de captação dos vínculos foi de 61,0% no ano de 2021. O principal avanço nesta abordagem é que foi possível vincular mais da metade das mortes maternas com nascimentos apenas com as técnicas de linkage empregadas. Há ainda de se pensar estratégias para melhorias destes achados, considerando que os bancos de dados não foram manipulados em sua essência. Apenas a idade foi ajustada e o estado foi selecionado para os dados de mortes maternas.

É possível também perceber que nos óbitos maternos não foram preenchidas informações acerca da gestação, o que poderia contribuir com a melhoria da vinculação. Ademais, pode-se considerar que algumas das mortes maternas também estejam vinculadas a mortes fetais, mas inicialmente este não é o foco da nossa estratégia de pesquisa.

LINKAGE DE SINAN E SIM

Para ilustração desta linkage, optamos por demonstrar com mortes por sífilis (17 casos em Pernambuco) sendo vinculado com o banco de dados sobre sífilis adquirida (3081 casos) do SINAN. Para a relação destes bancos de dados, apenas 4 variáveis foram identificadas como possíveis: sexo, idade, raça e código do município de residência.

A grande dificuldade é que o SINAN possui codificação e separação bem diferente da encontrada no SIM e SINASC. Por isso, foi necessário o ajuste das informações iniciais das variáveis que seriam vinculadas. Pelo tipo de dados, a estratégia mais conveniente era iniciar pelo linkage probabilístico.

Quadro 2 – Códigos empregados para realizar o linkage probabilístico do SINAN e SIM

```
pares_link_prob <- pares_blocagem |>
+   compare_pairs(by = c("sexo", "id", "racacor", "codmunres"),
+                 default_comparator = jaro_winkler(threshold = 0.9))

pares_link_prob

p3 <- pares_link_prob |>
+   score_simsum() |>
+   select_threshold(threshold = 3.6)

p3

p3 |>
+   as_tibble() |>
+   count(select)
# A tibble: 2 × 2
  select     n
    <lgl>  <int>
1 FALSE   4365
2 TRUE      12
```

Fonte: Pesquisa própria (2024)

A estratégia de linkage probabilístico é mais eficiente quando há variáveis com menos cobertura e qualidade. Neste caso de 17 mortes, 12 encontraram vinculações possíveis no banco do SINAN, o que representou 70,6% dos registros.

De forma análoga, uma estratégia de relação do SINASC e SINAN deve empregar a mesma estratégia. Há alguns óbices operacionais, como a falta de um identificador único para uma vinculação mais efetiva e uma padronização dos nomes de variáveis e codificação entre os SIS, além da necessidade de melhora no preenchimento de algumas variáveis.

CONSIDERAÇÕES FINAIS

A estratégia de linkage é uma excelente forma de interligar as informações de saúde de uma população, aprimorando as informações, de fácil acesso e baixo custo e pode ser incluída numa rotina de vigilância de desfechos em saúde.

AGRADECIMENTOS

Agradecemos à Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco pelo apoio financeito para realização do projeto.

REFERÊNCIAS

AGUIAR, F. P. *et al.* Análise da aplicação de uma rotina determinística para a identificação de gestações múltiplas no Sistema de Informações sobre Nascidos Vivos. **Epidemiologia e Serviços de Saúde**, v. 29, p. e2018454, 8 maio 2020.

ARECO, K. N. *et al.* Operational challenges in the use of structured secondary data for health research. **Frontiers in Public Health**, v. 9, p. 642163, 2021.

COELHO NETO, G. C.; CHIORO, A. Afinal, quantos Sistemas de Informação em Saúde de base nacional existem no Brasil? **Cadernos de Saúde Pública**, v. 37, n. 7, p. e00182119, 2021.

DRUMOND, E. de F. *et al.* Utilização de dados secundários do SIM, Sinasc e SIH na produção científica brasileira de 1990 a 2006. **Revista Brasileira de Estudos de População**, v. 26, p. 7–19, jun. 2009.

GARCIA, K. K. S.; MIRANDA, C. B. de; SOUSA, F. N. e F. de. Procedimentos para vinculação de dados da saúde: aplicações na vigilância em saúde. **Epidemiologia e Serviços de Saúde**, v. 31, p. e20211272, 10 out. 2022.

PETRUZALEK, D. READ. DBC: um pacote para importação de dados do datasus na linguagem R. **J. health inform**, v. 8, n. Supl 1, p. 601–605, 2016.

SALDANHA, R. de F.; BASTOS, R. R.; BARCELLOS, C. Microdatasus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). **Cadernos de Saúde Pública**, v. 35, n. 9, p. e00032419, 2019.

SILVA, N. P. da. **A utilização dos programas TABWIN e TABNET como ferramentas de apoio a disseminação das informações em saúde**. Rio de Janeiro: Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz, 2009.

ZHANG, D. *et al.* Information practices in data analytics for supporting public health surveillance. **Journal of the Association for Information Science and Technology**, v. 75, n. 1, p. 79-93, 2024.