ESTUDO DE CASOS SOBRE ANONIMIZAÇÃO DE DADOS NA LGPD

VERSÃO 1.0

BRASÍLIA/DF

NOVEMBRO DE 2023

Autoridade Nacional de Proteção de Dados

Diretor-Presidente

Waldemar Gonçalves Ortunho Júnior

Diretores

Arthur Pereira Sabbat

Joacil Basílio Rael

Miriam Wimmer

Equipe de Elaboração

Marcelo Santiago Guedes – Coordenador-Geral de Tecnologia e Pesquisa (CGTP)

Diego Carvalho Machado – Especialista (CGTP)

Albert França Josuá Costa – Especialista (CGTP)

SUMÁRIO

INTRODUÇÃO

Em busca de oferecer cenários e exemplos práticos de aplicação de técnicas de anonimização de dados, apresentam-se neste documento três estudos de caso. Tendo em vista a priorização da função didática, aos casos deu-se um recorte simplificado. Isso significa que se evitou descer a detalhes e a aspectos específicos que exigiriam abordagem de elementos não essenciais à implementação de técnicas de anonimização de dados. Importa salientar também que os estudos de caso estão em sintonia com os outros dois estudos técnicos sobre anonimização que analisam em profundidade o processo de anonimização de dados sob a perspectiva jurídico-regulatória e a computacional.

CASO 1

Dados agregados de localização – Supressão

A fim de adotar decisões informadas para combater emergência sanitária causada por epidemia de certa doença infectocontagiosa, a Secretaria Estadual de Saúde de um Estado-membro necessita de dados confiáveis a respeito da localização dos seus cidadãos. Os dados de localização são relevantes para que as autoridades estaduais do sistema de saúde pública possam identificar aglomerações de pessoas e, assim, orientarem-se para a implementação de medidas de prevenção, controle e fiscalização sanitárias de forma mais eficaz. A restrição ao ajuntamento das pessoas é medida importante para conter e diminuir o índice de contágio, além de identificar tendências de movimentação.

Com base na legislação instituidora de política pública de vigilância epidemiológica, discussões internas e consulta a especialistas, o governo estadual firmou acordo com os provedores de serviço de telefonia móvel A, B e C, para ter acesso a dados agregados de localização dos celulares dos respectivos usuários, com limites fixados à circunscrição territorial do Estado-membro e à duração da emergência sanitária.

Sendo assim, os provedores ou operadoras de telefonia móvel A, B e C compartilharam dados dos aparelhos de telefonia móvel conectados às Estações Rádio Base – ERBs. Cada aparelho de telefonia móvel envia para a ERB a que estão conectados a Identidade Internacional do Assinante Móvel (*Internation Mobile Subscriber Indentify* – IMSI) e a Identidade Internacional do Equipamento Móvel (*Internation Mobile Equipament Identify* – IMEI). Esses dados permitem que essas operadoras consigam identificar quais usuários estão conectados em quais ERBs em um determinado momento. Entretanto, para alcançar o objetivo da Secretaria Estadual de Saúde é necessário conhecer somente o quantitativo de usuários conectados em cada ERB em certo marco temporal.

Para resguardar a privacidade dos titulares de linhas móveis e atender ao interesse público, as operadoras de telefonia móvel, ao compartilharem os dados com a Secretaria de Saúde, aplicaram a técnica de supressão dos dados IMSI e IMEI, além de realizar a agregação do quantitativo de telefones móveis a fim de permitir o cálculo do índice de isolamento ou mapas de calor. Para tanto, consideraram-se: (i) o total de 21.641.000, o número de celulares somados os clientes das operadoras A, B e C no

território do Estado; e (ii) a localização a partir das antenas (Estações Rádio Base – ERBs) às quais os dispositivos móveis estavam conectados.

CASO 2

Dados clínicos para pesquisa acadêmica – Supressão e Pseudonimização

Em estudo de dados clínicos de pacientes conduzido por grupo de pesquisadores de determinado Hospital das Clínicas de uma universidade federal, os dados relacionados à pressão arterial de 100 pacientes foram coletados nos atendimentos realizados com intervalo de 7 (sete) dias. Os dados coletados estão dispostos na Tabela 1.

Tabela 1. Dados coletados

Nome Completo	CPF	Endereço	Gênero	Idade	Peso	PD 1	PS1	PD 2	PS 2
Johanne Mendonça	111.111.111-11	Rua Norte, 372 – Bairro A	M	51	113,30	13	9	15	7
Araci Coutinho Silva	222.222.222-22	Rua Leste, 122 – Bairro A	F	46	48,50	10	6	12	9
Marcela Antunes	333.333.333-33	Rua dos Cocos, 7 - Bairro B	F	37	97,44	10	7	11	7
Madruga Neves	444.444.444-44	Rua das Mangas, 22 - Bairro B	M	41	59,28	14	7	11	8
Florinda Neves	555.555.555-55	Rua das Mangas, 22 - Bairro B	F	58	54,30	11	7	11	7
Nilce Cavalcante	666.666.666-66	Rua Marte, 1 - Bairro C	F	57	110,33	15	6	12	8
José Francisco	777.777.777-77	Rua Vênus, 36 - Bairro C	M	73	58,55	18	10	17	10
Carmélia Andrade	888.888.888-88	Rua Vênus, 812 - Bairro C	F	56	54,42	12	7	12	7
Andreia Priscila	999.999.999-99	Rua Sol, 12 - Bairro C	F	35	109,38	17	10	16	9
...	...	...	...	...	...	...	...	...	...

Os pesquisadores submeteram esse conjunto de dados pessoais a processo de anonimização, tendo em vista que, conforme o desenho metodológico da pesquisa, a utilidade dos dados obtidos a partir da aplicação de certas técnicas de anonimização é preservada para os objetivos do estudo. Nesse sentido, foram aplicadas as técnicas expostas na Tabela 2. Cumpre ressaltar, ainda, que **os dados anonimizados serão**

mantidos em ambiente com controle de acesso e com pertinentes medidas de segurança previstas na política de segurança da informação do órgão de pesquisa.

Tabela 2 Técnicas utilizadas por Identificador.

Identificador	Técnica Utilizada	Descrição
Nome Completo	Supressão	Identificador direto é suprimido.
CPF	Pseudonimização	Substituição do CPF por um código único gerado.
Endereço	Supressão	O identificador é suprimido, pois não é útil para atender ao objetivo do tratamento.
Gênero		O processo de anonimização deve considerar a utilidade do dado para o tratamento desejado. No presente caso, os dados de gênero, peso, pressão diastólica 1, pressão sistólica 1, pressão diastólica 2 e pressão sistólica 2 estão correlacionados e essa correlação é útil para a finalidade da coleta de dados. Aplicação de técnicas de anonimização pode impactar na correlação dos dados e reduzir a utilidade deles.
Peso		O processo de anonimização deve considerar a utilidade do dado para o tratamento desejado. No presente caso, os dados de gênero, peso, pressão diastólica 1, pressão sistólica 1, pressão diastólica 2 e pressão sistólica 2 estão correlacionados e essa correlação é útil para a finalidade da coleta de dados. Aplicação de técnicas de anonimização pode impactar na correlação dos dados e reduzir a utilidade deles.
Pressão Diastólica 1		O processo de anonimização deve considerar a utilidade do dado para o tratamento desejado. No presente caso, os dados de gênero, peso, pressão diastólica 1, pressão sistólica 1, pressão diastólica 2 e pressão sistólica 2 estão correlacionados e essa correlação é útil para a finalidade da coleta de dados. Aplicação de técnicas de anonimização pode impactar na correlação dos dados e reduzir a utilidade deles.
Pressão Sistólica 1		O processo de anonimização deve considerar a utilidade do dado para o tratamento desejado. No presente caso, os dados de gênero, peso, pressão diastólica 1, pressão sistólica 1, pressão diastólica 2 e pressão sistólica 2 estão correlacionados e essa correlação é útil para a finalidade da coleta de dados. Aplicação de técnicas de anonimização pode impactar na correlação dos dados e reduzir a utilidade deles.
Pressão Diastólica 2		O processo de anonimização deve considerar a utilidade do dado para o tratamento desejado. No presente caso, os dados de gênero, peso, pressão diastólica 1, pressão sistólica 1, pressão diastólica 2 e pressão sistólica 2 estão correlacionados e essa correlação é útil para a finalidade da coleta de dados. Aplicação de técnicas de anonimização pode impactar na correlação dos dados e reduzir a utilidade deles.
Pressão Sistólica 2		O processo de anonimização deve considerar a utilidade do dado para o tratamento desejado. No presente caso, os dados de gênero, peso, pressão diastólica 1, pressão sistólica 1, pressão diastólica 2 e pressão sistólica 2 estão correlacionados e essa

		correlação é útil para a finalidade da coleta de dados. Aplicação de técnicas de anonimização pode impactar na correlação dos dados e reduzir a utilidade deles.

CASO 3

Compartilhamento de dados educacionais – Supressão, generalização, mascaramento, adição de ruídos e permutação

A Secretaria Municipal de Educação da cidade de Privacipolis precisa compartilhar os dados dos alunos matriculados com a Secretaria Municipal de Assistência Social com o objetivo da construção de relatórios sociais. Os dados estão dispostos na Tabela13.

Tabela13: Dados tratados.

Nome Completo	Matrícula	Idade	Endereço	Gênero	Renda Familiar (R$)
Johanne Mendonça	2023010	7	Rua Norte, 372 – Bairro A	M	2188,44
Araci Coutinho Silva	2023011	10	Rua Leste, 122 – Bairro A	F	2195,82
Marcela Antunes	2023020	8	Rua dos Cocos, 7 - Bairro B	F	1947,20
Madruga Neves	2023021	9	Rua das Mangas, 22 - Bairro B	M	2014,38
Florinda Neves	2023022	11	Rua das Mangas, 22 - Bairro B	F	1942,34
Nilce Cavalcante	2023030	12	Rua Marte, 1 - Bairro C	F	1856,08
José Francisco	2023031	12	Rua Vênus, 36 - Bairro C	M	1835,86
Carmélia Andrade	2023032	8	Rua Vênus, 812 - Bairro C	F	1989,66
Andreia Priscila	2023033	10	Rua Sol, 12 - Bairro C	F	2082,96
Bruno da Costa	2023040	13	Rua Mercúrio, 36 - Bairro C	M	1911,34

Incialmente, é preciso conhecer os dados tratados, o resumo dos dados está exposto na tabela abaixo.

Tabela24: Descrição dos dados.

Dado	Tipo	Dado Pessoal	Dado Pessoal Sensível	Identificador Direto	Descrição Estatística
Nome Completo	Qualitativo	S	N	S	Não Aplicável
Matrícula	Qualitativo	S	N	S	Não Aplicável
Idade	Quantitativo	S	N	N	Média: 10 Mediana: 10 Desvio-Padrão: 1,89
Endereço	Qualitativo	S	N	N	Não Aplicável
Gênero	Qualitativo	S	N	N	Moda: F

					Frequência M: 4/10 Frequência F: 6/10
Renda Familiar	Quantitativo	N	N	N	Média: R$ 1996,41 Mediana: R$ 1968,43 Desvio-Padrão: R$ 119,34 Mínimo: R$ 1835,85 Máximo: R$ 2195,82

Considerando o processo proposto no documento intitulado Estudo Técnico sobre Anonimização de Dados na LGPD: Processo de Anonimização Baseado em Risco e Técnicas de Anonimização – Uma Introdução Computacional, há 4 etapas essenciais para a gestão do risco de reidentificação.

- Determinar o Risco de Reidentificação Aceitável (RRA): É importante observar que a mensuração do risco de reidentificação é uma etapa que deve ser executada e gerenciada pelo agente de tratamento de acordo com o caso concreto, conforme sugerido no documento de Estudo Técnico sobre Anonimização de Dados na LGPD: Processo de Anonimização Baseado em Risco e Técnicas de Anonimização – Uma Introdução Computacional. No presente estudo de caso, nenhum dos dados tratados é considerado como sendo dado pessoal sensível e o compartilhamento dos dados é feito com outro órgão público por meios próprios. Entretanto, os dados são de crianças e adolescentes. De tal forma, o Risco de Reidentificação Aceitável (RRA) é definido em 0,35.
- Anonimizar os dados: A Tabela 35 apresenta as técnicas utilizadas em cada um dos dados tratados. Por sua vez, a Tabela 46 apresenta o conjunto de dados após a aplicação do conjunto de técnicas de anonimização.

Tabela 35: Técnicas utilizadas por Identificador.

Identificador	Técnica Utilizada	Descrição
Nome Completo	Supressão	Identificador direto que será suprimido, a matrícula será utilizada.
Matrícula	Mascaramento	Os dois primeiro e o último digito será substituído por *.
Idade	Generalização	Os dados serão agrupados por duas faixas etárias. 1ª_ ≤ 10 e 2ª >10
Endereço	Generalização	Os dados serão agrupados pelo bairro do endereço.
Gênero	Permutação	Os valores serão trocados entre os gêneros, porém mantendo a frequência de cada gênero e a moda do conjunto de dados.
Renda Familiar	Adição de Ruído e Generalização	Cada valor individual será deslocado um desvio-padrão à direita e posteriormente generalizado em duas faixas de renda: ≤ R$ 2.000,00 e > R$ 2.000,00

Tabela 46: Identificadores após aplicação do conjunto de técnicas de anonimização.

Matrícula	Idade	Endereço	Gênero	Renda Familiar (R$)
**2301*	≤ 10	Bairro A	F	> 2.000,00
**2301*	≤ 10	Bairro A	M	> 2.000,00
**2302*	≤ 10	Bairro B	F	> 2.000,00
**2302*	≤ 10	Bairro B	F	> 2.000,00

**2302*	> 10	Bairro B	M	> 2.000,00
**2303*	> 10	Bairro C	F	≤ 2.000,00
**2303*	> 10	Bairro C	M	≤ 2.000,00
**2303*	≤ 10	Bairro C	F	> 2.000,00
**2303*	≤ 10	Bairro C	M	> 2.000,00
**2304*	> 10	Bairro C	F	> 2.000,00

- Risco de Reidentificação Mensurado (RRM): O processo indica que após a aplicação do conjunto de técnicas de anonimização é necessário mensurar o risco de reidentificação utilizando alguma métrica contextual. Nesse estudo, optou-se por utilizar a K-Anonimização, métrica derivada da equivalência de classe. Conforme sugerido no processo, a métrica deve ser computada para cada um dos identificadores e os valores resultados ponderados para determinar o valor geral do risco mensurado de reidentificação (Tabela 57).

Tabela 57: Risco Mensurado de Reidentificação

Identificador	K-Anonimização por Classe do Identificador	K-Anonimização do Identificador (Média da K-Anonimização por Classe do Identificador)
Matrícula	**2301* = $\frac{1}{2}$ =0,50 **2302* = $\frac{1}{3}$ = 0,33 **2303* = $\frac{1}{4}$ = 0,25 **2304* = $\frac{1}{1}$ = 1,00	0,52
Idade	≤ 10 = $\frac{1}{6}$ = 0,16 > 10 = $\frac{1}{4}$ = 0,25	0,20
Endereço	Bairro A = $\frac{1}{2}$ = 0,50 Bairro B = $\frac{1}{3}$ = 0,33 Bairro C = $\frac{1}{5}$ = 0,20	0,34
Gênero	F = $\frac{1}{6}$ = 0,16 M = $\frac{1}{4}$ = 0,24	0,20
Renda Familiar (R$)	> 2.000,00 = $\frac{1}{8}$ = 0,12 ≤ 2.000,00 = $\frac{1}{2}$ = 0,50	0,31
Métrica Contextual (Média da K-Anonimização do Identificador)		0,31

No caso em estudo, não foram identificadas variáveis contextuais que impactem significativamente no risco de reidentificação, sendo o fator de ponderação definido em 1,00. Conforme proposto no processo, o Risco Mensurado de Reidentificação é o valor resultante da ponderação entre as variáveis contextuais e a métrica contextual, no exemplo 1,00*0,31 = 0,31.

O Risco de Reidentificação Mensurado calculado é de 0,31, enquanto o Risco de Reidentificação Aceitável é de 0,35. De tal forma, o conjunto de dados após a aplicação do conjunto de técnicas de anonimização tem um risco de reidentificação menor do que o risco aceitável. De acordo com o processo proposto, é necessário acompanhar o risco

mensurado de reidentificação para que ele sempre se mantenha abaixo do risco de reidentificação aceitável.