

Potencial no monitoramento em saúde: explorando os dados abertos ambulatoriais do SUS

Potential for health monitoring: Exploring SUS outpatient open data

Felipe Ferré^{1,2}, Amanda Oliveira Lyrio², Samara Helena de Carvalho², Laís Lessa Neiva Pantuzza², Jéssica Barreto Ribeiro dos Santos², Ana Carolina de Freitas Lopes², Luciene Fontes Schluckebier Bonan²

DOI: 10.1590/2358-28982025E110004P

RESUMO O Sistema Único de Saúde (SUS) disseminou bilhões de registros administrativos correspondentes a três décadas de existência. Na contramão de dados hospitalares e de notificação fragmentados por serem orientados ao serviço, e não ao usuário do SUS, os dados ambulatoriais apresentam dados vinculáveis ao identificador pseudonimizado, viabilizando acompanhar o itinerário terapêutico. Este trabalho apresenta, uma ferramenta automatizada da Sala Aberta de Inteligência em Saúde (Sabeis) de processamento, com software livre, de microdados fornecidos via estratégia TabWin/TabNet a partir do diretório de transferência de arquivos, sem aparato sofisticado de Big Data; e descreve os dados abertos do Sistema de Informação Ambulatorial, de 2008 a 2023, empregando recursos modestos de hardware por especialistas em saúde pública e informática em saúde. Foram processados 8.106.361.265 registros, correspondentes a 3.135 procedimentos, 16.407 diagnósticos e 51.875.308 usuários, segundo o arquivo correspondente à Autorização de Procedimentos Ambulatoriais e Alta Complexidade/Custo. Verificou-se a crescente qualidade do identificador pseudonimizado, sobretudo a partir de 2022, com 0,8% dos usuários com mais de um sexo, mais de um estado de residência, acima de oito procedimentos ou cinco diagnósticos. A presente abordagem demonstrou a potencialidade para o acompanhamento de políticas públicas, utilizando os dados ambulatoriais com identificador pseudonimizado do usuário do SUS.

PALAVRAS-CHAVE Ciência de dados. Sistemas de gerenciamento de base de dados. Sistema Único de Saúde. Documentação. Avaliação de tecnologias em saúde.

ABSTRACT The Brazilian Unified Health System (SUS) has disseminated billions of administrative records corresponding to three decades of existence. Unlike hospital and notification data, which are fragmented as they are service-oriented rather than user-oriented, outpatient data can be linked via a pseudonymized identifier, enabling the tracking of therapeutic pathways. This paper presents an automated microdata processing tool from the Open Health Intelligence Platform (SABEIS) using open-source software to handle microdata provided through the TabWin/TabNet strategy from the file transfer directory, without relying on sophisticated Big Data. It describes the open data from the SUS Ambulatory Information System from 2008 to 2023, using modest hardware resources by public health and health informatics specialists. A total of 8,106,361,265 records were processed, corresponding to 3,135 procedures, 16,407 diagnoses, and 51,875,308 users, according to the files of APAC-SIA High-cost ambulatory procedures and high-cost medicines. There was a noticeable improvement in the quality of the pseudonymized identifier, especially from 2022 onwards, with 0.8% of users recorded as having more than one sex, more than one state of residence, over eight procedures, or five diagnoses, demonstrating the potential for public policy monitoring and knowledge generation. This approach demonstrated the potential for monitoring public policies using outpatient data with a pseudonymized SUS user identifier.

KEYWORDS Data science. Database management systems. Unified Health System. Documentation. Technology assessment, health.

¹ Conselho Nacional de Secretários de Saúde (Conass) – Brasília (DF), Brasil.
labxss@gmail.com

² Ministério da Saúde (MS), Secretaria de Ciência, Tecnologia, Inovação e Complexo da Saúde (Sectics), Departamento de Gestão e Incorporação de Tecnologias em Saúde (DGITS) – Brasília (DF), Brasil.



Introdução

Apesar da abundância de registros e sistemas de informação com dados abertos, a gestão, o controle social e a comunidade acadêmica lidam, por três décadas, com dados fragmentados e incompletos. Isso se deve ao fato de que nem todas as políticas públicas do Sistema Único de Saúde (SUS) disseminaram, até a presente data, microdados consolidados de diferentes fontes. Destacam-se as notáveis ausências históricas de microdados abertos da Atenção Primária e da Assistência Farmacêutica, tanto em seus componentes básicos quanto estratégicos. Entretanto, iniciativas como a da Rede Interagencial de Informações para a Saúde (Ripsa) foram capazes de prover, ao longo de pelo menos 20 anos, capacidade analítica com indicadores com dados agregados que mostraram a rápida transformação das Redes de Atenção à Saúde (RAS), sobretudo de serviços sem internação¹.

A acertada estratégia de disponibilização de microdados, isto é, registros de contatos assistenciais e atividades administrativas sem agregação, tem sido crucial para valorizar os gestores e trabalhadores do SUS. As séries históricas que contemplam as primeiras decisões implantadas no SUS com as Normas Operacionais Básicas (NOB), políticas nacionais e demais atos normativos, quando integradas, viabilizam a avaliação longitudinal do itinerário terapêutico realizado. Além disso, o uso de dados abertos contribui para o monitoramento dos serviços oferecidos pelo SUS e para aprimorar a qualidade dos próprios dados^{2,3}.

Ainda que a transparência ativa da gestão pública apresente entraves e que a disseminação de microdados seja parcial⁴, o SUS construiu, com o TabNet e o TabWin, uma tecnologia bem-sucedida de tabulação e uso de dados não agregados. A tecnologia de disseminação implantada no decorrer das décadas de 1990 e 2000 se mostrou robusta e válida até o presente momento; e conta atualmente com uma massa crítica de sanitaristas letrados nessas ferramentas. Essas ferramentas

têm sido fundamentais na produção de uma grande parte dos indicadores e quantitativos para planos de saúde e relatórios de gestão⁵⁻⁷.

As ferramentas e dados oficiais, porém, não apresentam os microdados disponíveis no SUS das Ações e Serviços Públicos de Saúde (ASPS) diferentes de forma integrada, o que limita algumas análises. Por exemplo, os dados do Sistema de Informação Ambulatorial (SIA) incluem um identificador pseudonimizado do usuário do SUS, permitindo o acompanhamento do itinerário dentro do SUS. No entanto, a análise desses dados requer métodos sofisticados que envolvem técnicas capazes de lidar com bilhões de registros, cujo conhecimento geralmente não está disponível para conselheiros de saúde, sanitaristas, epidemiologistas ou profissionais que realizam Avaliação de Tecnologias de Saúde (ATS), em análises necessárias para que novos medicamentos, produtos e procedimentos sejam incorporados ao SUS⁸.

A lacuna de dados integrados e a indisponibilidade de métodos oficiais de curadoria obrigam cada núcleo de pesquisa e gestão ao desenvolvimento de ferramentas próprias para extração, transformação e carga (Extract, Transform and Load – ETL) de dados abertos produzidos via SIA, cuja consolidação de dados pode apresentar variações conforme o método aplicado^{5,9-16}.

O presente trabalho reúne as ciências ‘informática em saúde’ e a ‘sanitária’, cuja conciliação é estratégica para o País^{17,18}, focalizada nos insumos para tomada em decisão de saúde pública com dados de mundo real. Não foram detectados relatos de processamento completo e automatizado do SIA com *software* livre e aberto, apenas experiências de consolidação sem enfoque na integração de dados ambulatoriais ou com ferramentas proprietárias¹⁰.

Também não foram encontrados trabalhos com a base de dados abertos completa do SIA, problematizando o identificador pseudonimizado do usuário do SUS, lastreado no Cartão Nacional de Saúde (CNS), o qual é amplamente utilizado para tomada de decisão no SUS, embora exista ferramenta que aplica o

relacionamento determinístico de informações diagnósticas e de tratamento obtidas no SIA voltada para oncologia¹⁹. Entretanto, existe ampla literatura que relata o uso de pareamento determinístico-probabilístico com dados de acesso restrito por conterem informações de identificação do usuário do SUS, como nome completo, nome da mãe, documento de identificação, endereço com logradouro, ente outros^{13,14,20-23}, cujos microdados com identificador pseudonimizado derivado das ferramentas de vinculação não estão disponíveis para comparação.

Embora a quantidade de informações disponíveis no SUS cresça exponencialmente, ainda persistem lacunas que demandam atenção para subsidiar políticas públicas e assegurar um atendimento integral aos usuários do sistema. Portanto, o objetivo deste estudo é explorar os dados abertos do SIA e investigar as potenciais oportunidades de monitoramento longitudinal, adotando o identificador pseudonimizado como rastreamento do usuário do SUS, que permitem auxiliar, em especial, a tomada de decisão relacionada com incorporação, exclusão ou alteração de tecnologias de saúde no âmbito do SUS.

Material e métodos

Este estudo caracteriza-se como uma pesquisa exploratória e descritiva, fundamentada na análise de dados secundários oriundos de sistemas de informação do SUS, com enfoque na construção e estruturação de uma base integrada de dados a partir de múltiplos subsistemas.

Fonte e natureza dos dados

Os dados utilizados são de acesso público e foram obtidos a partir do repositório do Ministério da Saúde. Os arquivos de dados são disseminados, fragmentados por subsistema, 27 Unidades da Federação (UF) e data de competência com formato ano e mês em dois dígitos cada (AAMM), totalizando milhares de arquivos a serem processados.

Arquitetura e ambiente computacional

O processo de ETL foi realizado utilizando GNU Bash 5.2.15 para automação de rotinas; SQL (PostgreSQL 15.6, Ubuntu 23.10.1) para estruturação, transformação e carga dos dados; Wine 8.0.1 para emulação do descompactador oficial dbf2dbc.exe; dbview para extração dos dados DBF em formato CSV. O ambiente computacional consistiu em um minipc Lenovo ThinkCentre M75q Gen 2, com 32 GB de RAM, processador AMD Ryzen™ 5 PRO 5650GE ×12, e SSD Kingston SNV2S4000G (4TB). O código-fonte e o repositório de dados estão disponíveis sob as licenças General Public License (GPL 3.014) e Open Database License (ODbL).

Processamento e integração dos dados (ETL)

A consolidação dos dados seguiu as etapas conforme a metodologia Sala Aberta de Inteligência em Saúde (Sabeis)²⁴: 1) coleta automatizada, incluindo uma verificação do tamanho do arquivo (bytes) para assegurar a integridade do *download*; 2) descompactação e conversão dos arquivos do formato DBC para DBF, com posterior extração para CSV; 3) carga no Sistema Gerenciador de Banco de Dados (SGBD), como camada intermediária de dados (*staging*) brutos ou semibrutos; 4) validação da carga, verificando a correspondência do número de registros entre o arquivo original e os dados carregados no banco PostgreSQL, garantindo a consistência; 5) comparação da quantidade aprovada de procedimento com os dados oficiais disponibilizados pelo tabulador TabNet²⁵; 6) padronização de atributos conforme o dicionário de dados do SUS, mantendo uma estrutura unificada entre os subsistemas do sai; 7) carga das tabelas SQL orientadas a objetos no SGBD, utilizando o recurso Inherits para estruturar ‘tabelas-mãe’ por subsistema; 8) geração de *data marts*^{26,27}, respeitando as regras de negócio

para análise e relatórios; 9) adição do CNS no arquivo Produção Ambulatorial (PA) a partir de outros subsistemas, utilizando-se o código de autorização do procedimento, respectivo ao gestor estadual e o mês competência²⁸. O identificador do usuário é o mesmo ao longo dos diferentes subsistemas.

A Resolução nº 510, de 7 de abril de 2016, do Conselho Nacional de Saúde²⁹, prevê no art. 1º, inciso V, que pesquisas com ‘informações de acesso público’ não necessitam de avaliação por Comitê de Ética em Pesquisa (CEP).

Resultados

Foram processados dados do SIA de 2008 a 2023 (tabela 1) e realizada a qualificação do identificador pseudonimizado do usuário

(tabelas 2 a 4), avaliando sua trajetória do SUS (tabelas 2 e 5), cuja extração resultou em cerca de 1 terabyte em disco. Foram processados 44.637 arquivos de dados, contendo 8.106.361.265 registros, sendo que foi necessário enriquecer 5.024.137.874 registros do arquivo PA com os demais laudos para obter o identificador pseudonimizado.

Segundo a residência, todos os municípios e o Distrito Federal apresentaram, ao menos, um registro no SIA. Entretanto, é possível afirmar que nem todos os municípios ou regiões administrativas do Distrito Federal registraram procedimento no período avaliado para todas as políticas (tabela 1). A política vigente com menor número de municípios contemplados foi Acompanhamento Multiprofissional (AMP), com 1.050.

Tabela 1. Sumário dos dados administrativos produzidos via Sistema de Informação Ambulatorial de 2008 a 2023

Subsistema	Período	Arquivos originais	Mbytes FTP	Mbytes banco de dados**	Registros	Lista de Procedimentos***	Lista de Diagnósticos****	Municípios de residência
AB	01/2008-04/2017	544	438	897	211.252	1	-	1.425
ACF	08/2014-12/2023	2.945	14	29	315.631	1	-	5.378
AD	01/2008-12/2023	5.182	2.453	6.225	49.021.697	211	2.518	5.597
AM	01/2008-12/2023	5.131	13.115	35.958	287.923.699	608	660	5.591
AMP	03/2016-12/2023	541	3.821	6.570	64.900	2	-	1.050
AN	01/2008-10/2014	2.145	305	508	6.534.272	8	376	5.413
AQ	01/2008-12/2023	5.143	103.923	252.140	49.523.199	157	647	5.596
AR	01/2008-12/2023	4.683	137.080	563.504	3.615.964	39	719	5.595
ATD	08/2014-12/2023	3.047	51	322	12.391.833	6	-	5.551
BI	01/2008-12/2023	5.184	2.242	12.156	2.546.881.910	1.893	16.199	5.599

Tabela 1. Sumário dos dados administrativos produzidos via Sistema de Informação Ambulatorial de 2008 a 2023

Subsis- tema	Período	Arquivos originais	Mbytes FTP	Mbytes banco de dados**	Registros	Lista de Procedi- mentos***	Lista de Diagnósticos ****	Municípios de residência
PA	01/2008- 12/2023	5.447	24	62	5.024.137.874	3.135	16.184	5.599
PS	11/2012- 12/2023	3.557	762	1.380	122.214.893	21	3.858	5.064
SAD	04/2012- 10/2018	1.088	4	13	3.524.141	69	4.552	530
Total	01/2008- 12/2023	44.637	264.232	879.764	8.106.361.265	3.135	16.407	5.599*

Fonte: elaboração própria.

* Objetivando avaliar o acesso, optou-se por utilizar as Regiões Administrativas (RA) que constituem o Distrito Federal, logo, consideraram-se 5.569 municípios e 30 RA com registros no SUS (apesar de existirem 35 regiões desde o Decreto nº 38.094/2017).

** Carga com atributos selecionados. *** A lista da Renases (Relação Nacional de Ações e Serviços de Saúde) é mantida pelo Sistema de Gerenciamento da Tabela de Procedimentos, Medicamentos e Órteses, Próteses e Materiais Especiais do SUS (Sigtap), tendo sido contabilizados os códigos distintos. **** Lista de códigos distintos conforme a Classificação Internacional de Doenças, décima revisão (CID-10).

AB = Acompanhamento a Cirurgia Bariátrica; ACF = Confecção de Fístula Arteriovenosa; AD = Laudos Diversos; AM = Medicamentos; AMP = Acompanhamento Multiprofissional; AN = Nefrologia; AQ = Quimioterapia; AR = Radioterapia; ATD = Tratamento Dialítico; BI = Boletim Individual; PA = Produção Ambulatorial; PS = Psicossocial; SAD = Atenção Domiciliar. Sigtap – Sistema de Gerenciamento da Tabela de Procedimentos, Medicamentos e Órteses, Próteses e Materiais Especiais (OPM). Diagnóstico primário segundo a CID-10 (Classificação Internacional de Doenças). FTP protocolo de transferência de arquivos (File Transfer Protocol).

Quanto ao município de residência, todos os municípios e o Distrito Federal apresentaram pelo menos um registro de Laudos Diversos (AD), Medicamentos (AM), Quimioterapia (AQ), Boletim Individual (BI) e, consequentemente, PA. No SIA, não foram encontrados registros de Radioterapia (AR) para os municípios de Santa Isabel do Rio Negro-AM e Uiramutã-RR. Os municípios com menor cobertura relativa aos laudos do SIA foram Amaturá-AM, Nova Roma-GO, Serra Nova Dourada-MT, Nova Roma do Sul-RS, Putinga-RS e Nova Castilho-SP, jamais tendo registro de laudos para Acompanhamento a Cirurgia Bariátrica (AB), Confecção de Fístula Arteriovenosa (ACF), AMP, Nefrologia (AN), Tratamento Dialítico (ATD), Psicossocial (PS) e Atenção Domiciliar (SAD).

Com o identificador pseudonimizado, é possível avaliar o acesso a diferentes políticas de saúde pelo mesmo usuário, bem como

para o estadiamento no tratamento oncológico registrado via SIA³⁰. Na *tabela 2*, são apresentados números absolutos e por cem mil habitantes para o Brasil e as cinco regiões. Um exemplo é o acesso à política de tratamento oncológico (AQ ou AR) ofertada ao mesmo usuário de AM, ainda que não tenha ocorrido no mesmo período. É importante ressaltar que a maioria dos medicamentos registrados no SIA são do Componente Especializado da Assistência Farmacêutica (Ceaf). Entre 2008 e 2023, 266.500 usuários entre 2.666.836 (10,0%) de usuários de AQ acessaram AM, enquanto 170.633 usuários entre 1.666.261 (10,2%) de usuários de AR acessaram AM em algum momento. Os dados gerais podem situar a desigualdade das ASPS. Assumindo o número de registros como *proxy* de acesso, verificou-se que foi três vezes superior na região Sul em comparação à região Norte para AQ, e 1,7x superior ao observar a PA na totalidade.

Tabela 2. Usuários segundo subsistema do Sistema de Informação Ambulatorial de 2008 a 2023 em números absolutos (triângulo inferior em itálico) e número de usuários por cem mil habitantes por região do Brasil (2a) e utilização de mais de um tipo de serviço ambulatorial (2b)

	AB	ACF	AD	AM	AMP	AN	AQ	AR	ATD	BI	PA	PS	SAD
Tabela 2a													
Norte	1,1	72,6	4.582,8	1.532,1	1,8	85,0	626,9	386,6	164,1	56.927,2	17.560,9	2.139,7	23,5
Nordeste	2,1	108,8	6.219,6	3.105,7	4,2	123,6	1.090,9	673,8	232,5	38.917,8	28.031,8	2.512,8	44,1
Sudeste	14,7	104,2	6.592,4	6.366,5	10,5	157,7	1.447,4	938,4	252,4	64.636,1	26.230,9	2.250,0	83,3
Sul	59,8	108,7	7.521,2	4.953,5	13,1	159,5	1.924,5	1.162,6	276,8	92.123,0	29.955,7	3.062,5	40,3
Centro-Oeste	0,1	98,8	5.546,3	3.407,7	3,9	136,2	1.029,7	542,8	262,3	45.555,3	18.145,2	2.173,0	32,6
	AB	ACF	AD	AM	AMP	AN	AQ	AR	ATD	BI	PA	PS	SAD
Tabela 2b													
AB	31.757	0,1	11,7	12,0	0,0	0,1	1,3	1,0	0,1	156,0	183,0	4,6	0,1
ACF	16	208.654	708,2	908,5	16,0	119,2	27,1	15,0	1.126,5	804,1	1.202,3	15,0	1,6
AD	2.036	122.909	12.904.559	7.966,8	32,8	947,2	1.846,9	1.156,6	1.523,0	42.692,8	74.326,7	2.738,2	75,6
AM	2.075	157.673	1.382.627	9.362.825	30,5	1.068,5	1.535,6	983,2	1.793,7	27.003,1	53.946,4	2.349,8	87,3
AMP	0	2.771	5.701	5.298	16.034	0,3	3,4	2,1	21,6	68,6	92,4	1,3	0,1
AN	11	20.688	164.393	185.436	55	280.276	32,5	21,0	522,7	722,6	1.614,9	4,6	3,6
AQ	225	4.696	320.530	266.500	588	5.641	2.666.836	6.588,9	82,0	10.055,9	15.365,0	144,8	27,9
AR	168	2.605	200.720	170.633	368	3.651	1.143.500	1.666.261	47,0	6.471,8	9.600,0	100,8	19,8
ATD	26	195.497	264.319	311.303	3.757	90.709	14.239	8.150	485.605	1.715,8	2.798,1	32,4	5,4
BI	27.078	139.554	7.409.285	4.686.353	11.897	125.403	1.745.184	1.123.166	297.768	119.162.740	175.307,6	13.843,0	350,7
PA	31.757	208.654	12.899.319	9.362.329	16.034	280.258	2.666.584	1.666.074	485.600	30.424.434	51.875.308	7.809,5	259,3
PS	807	2.609	475.205	407.802	223	797	25.137	17.501	5.615	2.402.434	1.355.333	4.903.392	9,6
SAD	23	285	13.126	15.154	16	628	4.846	3.433	944	60.872	45.003	1.673	116.282

Fonte: elaboração própria.

AB = Acompanhamento a Cirurgia Bariátrica; ACF = Confeção de Fístula Arteriovenosa; AD = Laudos Diversos; AM = Medicamentos; AMP = Acompanhamento Multiprofissional; AN = Nefrologia; AQ = Quimioterapia; AR = Radioterapia; ATD = Tratamento Dialítico; BI = Boletim Individual; PA = Produção Ambulatorial; PS = Psicossocial; SAD = Atenção Domiciliar. População segundo o censo 2022.

As tabelas 3 e 4 contribuem para tipificar a qualidade do identificador pseudonimizado. A distribuição de sexo, raça/cor e idade apresenta consistência em relação ao que é praticado no SUS. Da mesma forma, é possível higienizar dado estrato desconsiderando identificadores pseudonimizados com número de registros elevado, o que pode indicar o uso do mesmo CNS para indivíduos diferentes. Adicionalmente, pode-se verificar o número de estados de residência, indiciando coincidência

do número utilizado por gestores diferentes; elevado número de procedimentos e diagnósticos estatisticamente impraticáveis para o mesmo sujeito; tempo de tratamento discrepante, entre outros quesitos. Ao avaliar usuários com ambos os sexos nos registros administrativos, mais de 2 estados de residência, 15 procedimentos ou 10 diagnósticos, apenas 3,05% se enquadram nessa situação, podendo, inclusive, ser retirados da análise como critério de exclusão.

Tabela 3. Qualidade do identificador pseudonimizado baseado no Cartão Nacional de Saúde relativo ao total de usuários com Produção Ambulatorial do Sistema de Informação Ambulatorial, entre 2008 e 2023, e percentual de registros a investigar por período, por região*

Variável	Usuários	Norte	Nordeste	Sudeste	Sul	Centro-Oeste
Total	51.875.308	3.000.390	15.093.470	21.985.990	8.810.592	2.903.022
Sexo**						
Apenas feminino	56,6%	56,3%	59,4%	55,0%	56,6%	53,8%
Apenas masculino	40,1%	38,6%	37,6%	41,4%	41,1%	41,6%
Ambos	3,4%	5,1%	3,0%	3,6%	2,3%	4,6%
Raça/cor***						
não identificado	35,5%	30,6%	41,7%	33,2%	33,1%	32,9%
branca	30,8%	5,1%	8,0%	40,2%	57,7%	23,4%
preta/parda	30,0%	58,7%	42,4%	25,2%	8,1%	38,8%
amarela	3,7%	5,5%	7,9%	1,5%	1,1%	4,5%
indígena	0,1%	0,1%	0,0%	0,0%	0,0%	0,4%
Idade***						
00-17	11,7%	14,5%	12,8%	10,2%	12,2%	12,0%
18-40	25,0%	27,4%	27,5%	22,6%	25,6%	25,1%
41-60	31,3%	31,6%	31,5%	31,5%	30,4%	31,9%
60+	31,8%	26,1%	27,8%	35,5%	31,5%	30,8%
desconhecida	0,3%	0,3%	0,3%	0,3%	0,3%	0,3%
Estados de residência						
1	98,9%	98,5%	99,0%	99,0%	98,9%	98,1%
2-14	1,1%	1,5%	1,0%	1,0%	1,1%	1,9%
Procedimentos						
1	47,7%	48,0%	49,6%	46,8%	45,8%	49,7%
2	19,3%	16,8%	19,8%	19,0%	19,8%	20,9%
3	9,3%	7,6%	9,4%	9,3%	9,9%	8,7%
4	5,9%	4,5%	5,7%	6,2%	6,0%	5,5%
5	3,9%	3,0%	3,8%	3,9%	4,1%	3,7%
6	2,9%	2,3%	2,9%	3,0%	2,9%	2,6%
7	2,2%	1,6%	2,2%	2,2%	2,1%	2,0%
8	1,7%	1,5%	1,6%	1,9%	1,7%	1,6%
9	1,4%	1,3%	1,1%	1,6%	1,4%	1,1%
10 a 220	5,8%	13,3%	3,9%	6,1%	6,3%	4,1%
Diagnósticos						
0	23,9%	26,9%	31,1%	17,4%	28,0%	19,8%
1	50,7%	48,6%	47,5%	53,3%	48,6%	56,1%
2	14,4%	13,6%	12,6%	16,0%	13,9%	14,8%
3	5,5%	5,2%	4,6%	6,4%	5,1%	5,1%
4	2,5%	2,3%	2,0%	3,0%	2,1%	2,0%
5	1,2%	1,3%	1,0%	1,5%	1,0%	0,9%
6	0,7%	0,7%	0,5%	0,8%	0,5%	0,4%
7 a 2.098	1,2%	1,4%	0,8%	1,6%	0,8%	0,8%

Tabela 3. Qualidade do identificador pseudonimizado baseado no Cartão Nacional de Saúde relativo ao total de usuários com Produção Ambulatorial do Sistema de Informação Ambulatorial, entre 2008 e 2023, e percentual de registros a investigar por período, por região*

Variável	Usuários	Norte	Nordeste	Sudeste	Sul	Centro-Oeste
Total	51.875.308	3.000.390	15.093.470	21.985.990	8.810.592	2.903.022
Investigar (usuários com ambos os sexos, acima de 2 estados, 15 procedimentos ou 10 diagnósticos)						
2008	8,5%	11,2%	7,3%	8,7%	8,4%	9,6%
2009 a 2015	5,8%	8,5%	5,3%	6,1%	5,2%	5,9%
2016 a 2021	6,4%	16,1%	4,5%	6,8%	5,4%	7,4%
2022	4,5%	15,3%	2,7%	4,5%	3,6%	4,6%
2023	3,1%	11,4%	1,8%	2,9%	2,8%	3,3%

Fonte: elaboração própria.

* Segundo Município de residência no primeiro atendimento. ** Sexo distintos registrados a cada atendimento. *** No primeiro atendimento

Na *tabela 4*, o valor de usuários a higienizar para tratamento oncológico, considerando critérios mais rigorosos de ambos os sexos, acima de 1 estado de residência, 8 procedimentos ou 6 diagnósticos, o número cai para 0,46% de usuários não unívocos, cujos registros devem ser higienizados adotando-se outros atributos (como idade e perfil de tratamento) ou excluídos da análise a depender do objetivo. A qualidade do identificador pseudonimizado cresce ao longo da evolução do CadSUS, com a conciliação do cadastro local em relação ao nacional, com a conciliação do (Master Patient Index – MPI) com o registro da Receita Federal realizado a

partir de 2016 e com a definição do Cadastro de Pessoa Física (CPF), como identificador único nacional, a partir de 2021. Ressalta-se a necessidade de higienização dos registros por usuários do SUS na região Norte em relação às demais, a qual permanece substancial em 2023 (11,4%), enquanto a menor encontra-se na região Nordeste com 1,8%, conforme os critérios adotados. Por sua vez, ao considerar políticas específicas, a exemplo da oncologia, o contingente a higienizar da região Norte é de 1,5% dos casos, enquanto no Sudeste é de 0,3% em 2023.

Tabela 4. Qualidade do identificador pseudonimizado baseado no Cartão Nacional de Saúde relativo ao total de usuários de oncologia (quimioterapia – AQ e radiologia – AR) com Produção Ambulatorial do Sistema de Informação Ambulatorial, entre 2008 e 2023, e percentual de registros a investigar por período, por região*

Variável	Usuário	Norte	Nordeste	Sudeste	Sul	Centro-Oeste
Total	3.168.693	135.069	701.393	1.455.544	672.032	204.655
Sexo**						
Apenas feminino	53,7%	57,5%	56,1%	52,7%	52,8%	53,5%
Apenas masculino	45,7%	41,2%	43,3%	46,8%	46,7%	45,8%
Ambos	0,6%	1,4%	0,6%	0,5%	0,5%	0,6%
Raça/cor***						
Branca	46,7%	8,7%	17,6%	52,6%	77,9%	27,0%
Preta/parda	33,8%	54,5%	56,0%	31,8%	8,8%	40,1%

Tabela 4. Qualidade do identificador pseudonimizado baseado no Cartão Nacional de Saúde relativo ao total de usuários de oncologia (quimioterapia – AQ e radiologia – AR) com Produção Ambulatorial do Sistema de Informação Ambulatorial, entre 2008 e 2023, e percentual de registros a investigar por período, por região*

Variável	Usuário	Norte	Nordeste	Sudeste	Sul	Centro-Oeste
Total	3.168.693	135.069	701.393	1.455.544	672.032	204.655
Não identificado	17,5%	35,8%	20,7%	14,5%	12,6%	31,9%
Amarela	2,0%	0,8%	5,7%	1,1%	0,7%	0,8%
Indígena	0,0%	0,2%	0,0%	0,0%	0,0%	0,1%
Idade***						
00-17	2,6%	5,0%	3,3%	2,2%	1,9%	3,3%
18-40	8,9%	13,8%	10,2%	7,9%	8,3%	10,5%
41-60	35,5%	38,6%	35,4%	34,9%	35,8%	37,6%
60+	52,8%	42,5%	50,9%	54,9%	53,8%	48,4%
Desconhecido	0,1%	0,2%	0,1%	0,1%	0,1%	0,2%
Estados de residência						
1	99,2%	96,0%	99,1%	99,6%	99,4%	98,5%
2-4	0,8%	4,0%	0,9%	0,4%	0,6%	1,5%
Procedimentos						
1	38,2%	37,9%	37,0%	38,2%	39,1%	39,7%
2	17,8%	17,2%	18,1%	17,6%	17,9%	18,6%
3	9,3%	11,7%	10,1%	9,1%	8,2%	9,4%
4	8,5%	9,8%	8,0%	9,2%	7,1%	8,5%
5	10,8%	10,1%	10,8%	11,1%	10,5%	9,5%
6	8,3%	7,4%	8,2%	8,1%	9,2%	7,8%
7	4,1%	3,5%	4,7%	3,8%	4,4%	3,8%
8	1,7%	1,3%	1,9%	1,6%	2,0%	1,6%
9-18	1,3%	0,9%	1,3%	1,2%	1,5%	1,2%
Diagnósticos						
1	70,5%	68,1%	69,8%	71,2%	70,3%	69,8%
2	24,0%	26,4%	25,2%	23,2%	23,6%	25,1%
3	4,5%	4,5%	4,2%	4,5%	4,9%	4,3%
4	0,8%	0,8%	0,6%	0,9%	1,0%	0,7%
5	0,2%	0,1%	0,1%	0,2%	0,2%	0,1%
6	0,0%	0,0%	0,0%	0,1%	0,0%	0,0%
7+	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
Investigar (usuários com ambos os sexos, acima de 1 estado, 8 procedimentos ou 6 diagnósticos)						
2008	3,8%	12,7%	3,6%	3,2%	3,5%	4,4%
2009 a 2015	3,2%	5,4%	3,6%	2,7%	3,2%	3,9%
2016 a 2021	2,4%	6,4%	2,4%	1,8%	2,5%	3,1%
2022	0,8%	3,8%	0,8%	0,5%	0,6%	1,1%
2023	0,5%	1,5%	0,4%	0,3%	0,4%	0,6%

Fonte: elaboração própria.

* Segundo Município de residência no primeiro atendimento. ** Sexo distintos registrados a cada atendimento. *** No primeiro atendimento.

A transparência oferecida pelos dados abertos e sua organização em repositórios, ou lagos de dados, possibilita explorar como algumas políticas são difundidas e realizar análises de situação exploratória e direcionadas por dados (*data-driven analysis*). Por exemplo, como observado na *tabela 5*, a hemodiálise apresentou 466.918 usuários, cujo custo médio, segundo números

oficiais sem deflacionamento, foi de R\$ 81.945,25, correspondendo a um tratamento longo, cuja mediana, em dias, foi 818 meses. Adicionalmente, destaca-se a reabilitação do desenvolvimento neuropsicomotor, cuja mediana de idade é 6 e o valor médio por usuário foi de 6.820,61, com 426 procedimentos por usuário, considerando uma média simples.

Tabela 5. Caracterização dos procedimentos realizados mais de três vezes por usuário com maior número de usuários do SUS, entre 2008 e 2023

Procedimento	Usuários	Registros por usuário	Quantidade por usuário	Valor por usuário (R\$)	Dias tratamento (mediana)	Idade (mediana)
Hemodiálise (máximo 3 sessões por semana)	466.918	36	445	81.945,25	818	56
Dosagem de ureia	495.295	33	63	116,26	849	56
Dosagem de potássio	482.057	34	34	63,05	848	56
Atendimento / acompanhamento de paciente em reabilitação do desenvolvimento neuropsicomotor	103.736	155	426	6.820,61	605	6
Dosagem de transaminase glutâmico-pirúvica (tgp)	477.795	33	33	67,33	849	56
Dosagem de fósforo	448.613	36	36	65,99	820	56
Dosagem de cálcio	451.827	35	35	65,54	847	56
Formoterol 12 mcg + budesonida 400 mcg pó inalante (por frasco de 60 doses)	457.793	33	35		1.034	59
Atendimento fisioterapêutico nas alterações motoras	512.731	27	69	324,80	147	50
Atorvastatina 20 mg (por comprimido)	461.091	28	1.221		879	62
Alfaepoetina 4.000 ui injetável (por frasco-ampola)	491.314	25	279		848	59
Dosagem de creatinina	495.165	19	19	35,97	819	56
Olanzapina 10 mg (por comprimido)	235.140	38	1.554	1.423,08	1.153	39
Hematócrito	339.472	26	26	40,33	849	56
Dosagem de hemoglobina	336.777	26	26	40,21	849	56
Hemograma completo	480.839	17	17	70,27	969	55
Consulta médica em atenção especializada	789.478	10	11	115,88	666	53
Formoterol 12 mcg + budesonida 400 mcg (por capsula inalante)	295.265	27	1.311		941	61
Dosagem de glicose	352.682	21	21	38,41	939	57

Tabela 5. Caracterização dos procedimentos realizados mais de três vezes por usuário com maior número de usuários do SUS, entre 2008 e 2023

Procedimento	Usuários	Registros por usuário	Quantidade por usuário	Valor por usuário (R\$)	Dias tratamento (mediana)	Idade (mediana)
Azatioprina 50 mg (por comprimido)	191.635	35	2.622		1.095	39
Hormonoterapia do adenocarcinoma de próstata avançado - 1ª linha	270.434	25	25	7.491,04	697	72
Atorvastatina 40 mg (por comprimido)	321.568	21	751		606	64
Dosagem de sódio	247.959	27	27	49,72	757	57
Sacarato de hidróxido férrico 100 mg injetável (por frasco de 5 ml)	345.532	19	82	355,58	817	57
Hormonoterapia do carcinoma de mama em estágio ii	177.361	36	36	2.883,72	1.154	57
Leflunomida 20 mg (por comprimido)	191.217	33	994	1.241,15	971	55
Tacrolimo 1 mg (por capsula)	115.763	51	6.692	229,86	1.549	46
Timolol 5,0 mg/ml solução oftálmica (por frasco de 5 ml)	219.855	25	25		909	65
Risperidona 2 mg (por comprimido)	193.391	28	1.668	132,41	789	38
Dosagem de fosfatase alcalina	365.816	15	15	29,39	1.031	55

Fonte: elaboração própria.

Nota: valores não deflacionados. Medicamentos não apresentam valor mantido no Sistema de Gerenciamento da Tabela de Procedimentos, Medicamentos e Órteses, Próteses e Materiais Especiais do SUS (Sigtap) e, portanto, não são registrados no SIA.

Discussão

O estudo destacou a capacidade de gerenciar e aprimorar dados abertos sem a presença de profissionais especializados em administração de servidores, mas sim com especialistas em saúde pública que possuem conhecimentos substanciais em tecnologia da informação, assumindo uma função próxima à de engenheiros de análise de dados em saúde²⁷.

O SUS apresentou no SIA 3.135 procedimentos ambulatoriais, segundo o Sistema de Gerenciamento da Tabela de Procedimentos, Medicamentos e Órteses, Próteses e Materiais Especiais (Sigtap), realizados sob 16.407 diagnósticos primários. Os dados abertos não são fornecidos orientados ao usuário, mas fragmentados à respectiva ASPS. Logo, quando ocorre a opção administrativa em descontinuar

no SIA uma dada ASPS, o usuário pode deixar de ser monitorado de forma integrada às demais políticas ambulatoriais e especializadas, como demonstrado na *tabela 1* com os dados de laudos de AN, a qual contém apenas dados de 2008 a 2014.

Ao avaliar a ausência de procedimentos realizados segundo o município de residência, os dados podem revelar vazios assistenciais, porém, podem indiciar uma cultura de cadastramento fora do domicílio para a obtenção do acesso, revelando fragilidades no Planejamento Regional Integrado (PRI) e conflitos federativos e, portanto, na pactuação entre municípios, uma vez que a resolutividade deve ocorrer na região de saúde, e não totalmente no município, não podendo ser cerceado o acesso para quem não habita no mesmo município de atendimento^{31,32}.

Embora os dados abertos sejam amplamente utilizados na produção acadêmica e pela gestão do SUS, o uso do CNS é questionado em fóruns em função das possíveis inconsistências, por exemplo, duplicidades relativas à disponibilidade de sistema de informação local de produção do cartão dos municípios e estados que geram um número próprio para seu cidadão sem sincronicidade com a base nacional mantida pelo Ministério da Saúde³³.

A avaliação da qualidade de identificadores pseudonimizados é fundamental, conforme ilustrado pelas *tabelas 3 e 4*. A análise permite a identificação e exclusão de registros cujo identificador contemple critérios de inconsistência, ante a possibilidade do uso do mesmo número de identificação entre diferentes gestores para pessoas diferentes, inferida pelo número de estados de residência, sexos diferentes e idades calculadas em relação ao primeiro registro e a data de atendimento ou tempos de tratamento discrepantes, bem como a presença de um número estatisticamente inviável de procedimentos e diagnósticos para um único sujeito, podem ser utilizados para refinar os dados. Por exemplo, apenas 3,05% dos registros apresentaram inconsistências como a presença de ambos os sexos, mais de 2 estados de residência, 15 ou mais procedimentos, ou 10 ou mais diagnósticos, validando a aplicação desses critérios de exclusão para garantir a integridade da análise.

Dessa forma, o presente trabalho demonstrou que a consolidação nacional dos dados e a oferta do identificador pseudonimizado podem ser trabalhadas de modo a constituir estratos populacionais segundo a política de saúde, configurando, inclusive, coortes conforme o tratamento ou diagnóstico.

É importante destacar o conhecimento aplicado por informatas em saúde utilizando simplicidade na codificação e recursos de *hardware* domésticos. Uma das dificuldades para o processamento de dados ambulatoriais é o manejo dos arquivos originais em formato DBC dotados de milhões de registros. O descompactador oficial fornecido pelo Ministério

da Saúde, o dbf2dbc.exe, foi o único viável para processar todos os arquivos de dados utilizando recursos computacionais domésticos sem falhas. Não se obteve consistência ao utilizar ferramentas alternativas em arquivos de estados com maior número de registros, como São Paulo e Minas Gerais, particularmente após 2019, ao utilizar alternativas para linguagens de programação estatísticas e interpretadas como projeto R e Python em computador doméstico^{9,34}.

O Bash é também uma linguagem interpretada, isto é, opera como interpretador de comandos, executando instruções fornecidas pelo sistema operacional. Entretanto, o Bash é projetado principalmente para uso interativo, o que significa que é otimizado para a execução imediata de comandos de sistema, sem a necessidade de compilação ou estruturação complexa de código. A vantagem do processamento em linguagem Bash está no enfoque de a ETL ser projetada para a automação de tarefas em sistemas Unix e Linux, com a execução de comandos do sistema operacional, manipulação de arquivos, processamento de texto e criação de *scripts* de terminal (*shell*).

O processamento foi realizado sem adotar ferramentas específicas para *big data*, evitando plataformas de ETL visual com funcionalidades de ‘arrastar e soltar’. Em vez disso, foram utilizadas técnicas de ingestão em Bash e operações SQL, seguindo boas práticas de desenvolvimento com *software* livre. Como resultado, a presente abordagem é otimizada para operações comuns ao ambiente de linha de comando ao invés de soluções de ETL como do ecossistema Pentaho ou Informática PowerCenter®, incluindo acesso direto e eficiente aos recursos do sistema operacional Unix/Linux, o que significa que muitas operações podem ser realizadas sem a sobrecarga associada à inicialização de uma máquina virtual (como no caso do Python) ou ao processamento de dados complexos (como no caso do R). Destaca-se, adicionalmente, a baixa sobrecarga de memória e processamento devido à natureza direta e enfoque em tarefas simples do sistema operacional³⁵⁻³⁸.

Fragilidades no repasse financeiro podem criar iniquidades no acesso ao ATD em diferentes territórios do Brasil e dificuldades na avaliação quando os dados não constam no sistema de informação, seja para procedimentos, seja o valor gasto com medicamento³⁹. Logo, lacunas na disseminação de dados oficiais vinculados ao sistema de informação de origem implicam, por exemplo, a coleta discrepante dos montantes gastos a partir de fontes diferentes, cujas estimativas podem variar conforme o método do estudo farmacoeconômico.

Conclusões

O presente trabalho demonstrou a importância da uniformização dos microdados abertos em estratégia de disseminação integrada de modo a avançar em análises longitudinais com dados administrativos, sobretudo por serem uma fonte comum e disponível para o debate amplo do Estado com a sociedade civil. Infelizmente, os dados hospitalares, de notificação de agravo, nascidos vivos, óbitos, imunização, existentes na mesma estratégia TabWin/TabNet, não são disponibilizados de forma integrada e aberta, pois, apesar de os registros serem individualizados por contato assistencial ou usuário, inviabilizam análise longitudinal do itinerário no SUS. Ainda assim, tais microdados reunidos sistematicamente e organizados em lagos de dados apresentam grande potencial para estudos ecológicos por município com sistemas de indicadores e salas de situação⁴⁰⁻⁴³.

Demonstrou-se a importância de a qualidade da ETL ser aferida a cada etapa e de sua resultante ser explicitada nos estudos envolvendo grandes volumes de dados disseminados. O presente ensejo recomenda a prática de ciência aberta com disponibilização do código-fonte, não apenas para a reprodutibilidade, mas, sobretudo, diante do recurso das redes neurais de processamento profundo e o inevitável uso das Inteligências Artificiais (IA) generativas. A qualidade explícita do

tratamento de dados pode ser critério para reduzir alucinações de IA e asserções espúrias, não criando outros propagadores de desinformação.

Espera-se que a transformação digital em saúde disponibilize lagos de dados integrados para que pesquisadores e gestores adotem os conhecimentos da informática em saúde com uma base com e não incorram na fragmentação de dados por serviço de gestão, atenção ou vigilância conduzida antes de haver a Política Nacional de Informação e Informática (PNIIS), mas que promova dados orientados ao usuário do SUS conforme preconizado nas pactuações tripartite⁴⁴⁻⁴⁶.

Colaboradores

Ferré F (0000-0001-9879-4782)* contribuiu para conceituação, investigação, curadoria de dados, análise formal, metodologia e escrita do manuscrito. Lyrio AO (0000-0001-7740-2524)* contribuiu para conceituação, investigação, curadoria de dados, análise formal, investigação, metodologia e escrita do manuscrito. Carvalho SH (0009-0005-7716-8164)* contribuiu para conceituação, investigação, curadoria de dados, análise formal, aquisição de financiamento, investigação, metodologia, administração do projeto, provimento de recursos, provimento de *software*, supervisão, validação, visualização, escrita, revisão e edição do manuscrito. Pantuzza LLN (0000-0001-9831-8807)* contribuiu para conceituação, curadoria de dados, escrita, revisão e edição do manuscrito. Santos JBR (0000-0002-5528-0658)* contribuiu para conceituação, curadoria de dados, escrita, revisão e edição do manuscrito. Lopes ACF (0000-0001-5806-5155)* contribuiu para conceituação, administração do projeto, provimento de recursos, supervisão, validação e escrita, revisão e edição do manuscrito. Bonan LFS (0000-0003-2857-2927)* contribuiu para administração do projeto, provimento de recursos e supervisão do manuscrito. ■

*Orcid (Open Researcher and Contributor ID).

Referências

1. Pinto LF, Freitas MPS, Figueiredo AWS. Sistemas Nacionais de Informação e levantamentos populacionais: algumas contribuições do Ministério da Saúde e do IBGE para a análise das capitais brasileiras nos últimos 30 anos. *Ciênc saúde coletiva*. 2018;23(6):1859-70. DOI: <https://doi.org/10.1590/1413-81232018236.05072018>
2. Fernandes FECV, Leal IS, Andrade JDA. Percepção dos profissionais da atenção primária à saúde sobre o sistema de informação ambulatorial. *Rev Enferm Atenção Saúde*. 2017;6(2):77-92. DOI: <https://doi.org/10.18554/reas.v6i2.1673>
3. Silva AR, Oliveira TM, Lima CF, et al. Sistemas de informação como instrumento para tomada de decisão em saúde: revisão integrativa. *Rev enferm UFPE on line*. 10(9):3455-62. DOI: <https://doi.org/10.5205/1981-8963-v10i9a11428p3455-3462-2016>.
4. Maduro-Abreu A, Litre G, Santos L, et al. Transparência da informação pública no Brasil: uma análise da acessibilidade de Big Data para o estudo das interfaces entre mudanças climáticas, mudanças produtivas e saúde. *RECIIS*. 2020;14(1):111-25. DOI: <https://doi.org/10.29397/reciis.v14i1.1690>
5. Silva NP. A utilização dos programas TABWIN e TABNET como ferramentas de apoio à disseminação das informações em saúde [dissertação]. Rio de Janeiro: Escola Nacional de Saúde Pública Sergio Arouca, Fundação Oswaldo Cruz; 2009.
6. Leandro BBS, Rezende FAV, Pinto JMC. Informações e registros em saúde e seus usos no SUS. Rio de Janeiro: Editora FIOCRUZ; 2020.
7. Garcia PT, Reis RS. Gestão pública em saúde: o plano de saúde como ferramenta de gestão. São Luís: EDUFMA; 2016.
8. Ferré F. Infoestrutura para apoio à decisão estratégica no SUS. In: Santos AO, Lopes LT, organizadoras. Reflexões e futuro. Brasília, DF: Conass; 2021. p. 114-27. (Coleção Covid-19; v. 6).
9. Saldanha RF, Bastos RR, Barcellos C. Microdatasus: pacote para download e pré-processamento de microdados do Departamento de Informática do SUS (DATASUS). *Cad Saúde Pública*. 2019;35(9):e00032419. DOI: <https://doi.org/10.1590/0102-311X00032419>
10. Santos RS, Santos RS, Gutierrez MA. MINERSUS Ambiente computacional para extração de informações para a gestão da saúde pública por meio da mineração dos dados do SUS. *Rev Bras Eng Biomed*. 2008;24(2):77-90. DOI: <https://doi.org/10.4322/rbeb.2012.050>
11. Franceschini PM, Porto JB, Kunst R. Ferramenta de Visualização de Dados Públicos da Saúde Disponibilizados pelo DATASUS. In: International Conference on Information Resources Management; 2021; [local desconhecido]: AIS Electronic Library; 2021.
12. Barbosa MN. Possibilidades e limitações de uso das bases de dados do DATASUS no controle externo de políticas públicas de saúde no Brasil [trabalho de conclusão de curso]. Brasília, DF: Instituto Cerzedello Corrêa, Escola Superior do Tribunal de Contas da União; 2019.
13. Cherchiglia ML, Guerra Júnior AA, Andrade EIG, et al. A construção da base de dados nacional em terapia renal substitutiva (TRS) centrada no indivíduo: aplicação do método de linkage determinístico-probabilístico. *Rev Bras Estud Popul*. 2007;24(1):163-67. DOI: <https://doi.org/10.1590/S0102-30982007000100010>
14. Barreto ML, Ichihara MY, Pescarini JM, et al. Cohort Profile: The 100 Million Brazilian Cohort. *Int J Epidemiol*. 2022;51(2):e27-e38. DOI: <https://doi.org/10.1093/ije/dyab213>
15. Moura L de, Prestes IV, Duncan BB, et al. Construção de base de dados nacional de pacientes em tratamento dialítico no Sistema Único de Saúde, 2000-2012. *Epidemiol Serv Saude*. 2014;23(2):227-38. DOI: <https://doi.org/10.5123/S1679-49742014000200004>

16. Moya J, Risi Junior JB, Martinello A. Salas de situação em saúde: compartilhando as experiências do Brasil. Brasília, DF: Organização Pan-Americana da Saúde; Ministério da Saúde; 2010.
17. Fornazin M, Joia LA. Articulando perspectivas teóricas para analisar a informática em saúde no Brasil. *Saude Soc.* 2015;24:46-60. DOI: <https://doi.org/10.1590/S0104-12902015000100004>
18. Giannotti EM, Fonseca F, Panitz LM. Sistemas de Informação da Atenção à Saúde: contextos históricos, avanços e perspectivas no SUS. Brasília, DF: Cidade Gráfica e Editora Ltda; 2015.
19. Atty ATM, Jardim BC, Dias MBK, et al. PAINEL-Onco: uma Ferramenta de Gestão. *Rev Bras Cancerol.* 2020;66(2); DOI: <https://doi.org/10.32635/2176-9745.rbc.2020v66n2.827>
20. Camargo Jr. KR, Coeli CM. Reclink: aplicativo para o relacionamento de bases de dados, implementando o método probabilistic record linkage. *Cad Saúde Pública.* 2000;16(2):439-47. DOI: <https://doi.org/10.1590/S0102-311X2000000200014>
21. Ali MS, Ichihara MY, Lopes LC, et al. Administrative Data Linkage in Brazil: Potentials for Health Technology Assessment. *Front Pharmacol.* 2019;10:984. DOI: <https://doi.org/10.3389/fphar.2019.00984>
22. Guerra Junior AA, Pereira RG, Gurgel EI, et al. Building the National Database of Health Centred on the Individual: Administrative and Epidemiological Record Linkage-Brazil, 2000-2015. *Int J Popul Data Sci.* 2018;3(1):446. DOI: <https://doi.org/10.23889/ijpds.v3i1.446>
23. Tomazelli JG, Girianelli VR, Silva GA. Estratégias usadas no relacionamento entre Sistemas de Informações em Saúde para seguimento das mulheres com mamografias suspeitas no Sistema Único de Saúde. *Rev Bras Epidemiol.* 2018;21:e180015. DOI: <https://doi.org/10.1590/1980-549720180015>
24. Ferré F, Oliveira G, Queiroz M, et al. Sala de Situação aberta com dados administrativos para gestão de Protocolos Clínicos e Diretrizes Terapêuticas de tecnologias providas pelo SUS. In: 20º Simpósio Brasileiro de Computação Aplicada à Saúde SBC; 2020 set 15-18; Porto Alegre. Porto Alegre: Sociedade Brasileira de Computação; 2020. p. 392-403.
25. TabNet [Internet]. Brasília, DF: DATASUS. c2008 [acesso em 2022 out 28]. Informações de saúde. Produção Ambulatorial do SUS por local de atendimento. Disponível em: <http://tabnet.datasus.gov.br/cgi/defthtohtm.exe?sia/cnv/qauf.def>
26. Elmasri R, Navathe SB. Sistemas de banco de dados. 6ª ed. São Paulo: Pearson Addison Wesley; 2011.
27. Khan B, Jan S, Khan W, et al. An overview of ETL techniques, tools, processes and evaluations in data warehousing. *J Big Data.* 2024;6. DOI: <http://dx.doi.org/10.32604/jbd.2023.046223>
28. Franco TB. Trabalho, cuidado e transição tecnológica na saúde: um olhar a partir do sistema cartão nacional de saúde. Porto Alegre: Editora Rede Unida; 2021 (Série Micropolítica do Trabalho e o Cuidado em Saúde).
29. Conselho Nacional de Saúde (BR). Resolução nº 510, de 7 de abril de 2016. Dispõe sobre as normas aplicáveis a pesquisas em Ciências Humanas e Sociais. *Diário Oficial da União, Brasília, DF.* 2016 maio 24; Edição 98; Seção I:44-46.
30. Atty ATM, Tomazelli JG, Dias MBK. Análise Exploratória das Informações sobre Estadiamento nas Autorizações de Procedimentos de Alta Complexidade no Brasil e Regiões no Período 2010-2014. *Rev Bras Cancerol.* 2019;63(4):257-64. DOI: <https://doi.org/10.32635/2176-9745.RBC.2017v63n4.126>
31. Medeiros CRG, Saldanha OMFL, Grave MTQ, et al. Planejamento regional integrado: a governança em região de pequenos municípios. *Saude Soc.* 2017;26(1):129-40. DOI: <https://doi.org/10.1590/S0104-12902017162817>
32. Lima LD, Viana ALD, Machado CV, et al. Regionalização e acesso à saúde nos estados brasileiros: condi-

- cionantes históricos e político-institucionais. *Ciênc saúde coletiva*. 2012;17(11):2881-92. DOI: <https://doi.org/10.1590/S1413-81232012001100005>
33. Cunha RE. Cartão Nacional de Saúde: os desafios da concepção e implantação de um sistema nacional de captura de informações de atendimento em saúde. *Ciênc saúde coletiva*. 2002;7(4):869-78. DOI: <https://doi.org/10.1590/S1413-81232002000400018>
 34. Petruzalek D. read.dbc: Read Data Stored in DBC (Compressed DBF) Files [Internet]. New York: Datacamp; 2016 [acesso em 2022 out 28]. Disponível em: <https://cran.r-project.org/web/packages/read.dbc/index.html>
 35. Park H, Lee S, Gim G, et al. Dataverse: Open-Source ETL (Extract, Transform, Load) Pipeline for Large Language Models. *arXiv [csCL]*. 2024. DOI: <https://doi.org/10.48550/arXiv.2403.19340>
 36. Ernest A, Mensah E, Gilbert A. Qualitative assessment of compiled, interpreted and hybrid programming languages. *Comm App Electronics*. 2017;7:8-13. DOI: <https://doi.org/10.5120/cae201765268>
 37. Sanner MF. Python: a programming language for software integration and development. *J Mol Graph Model*. 1999;17(1):57-61.
 38. Prechelt L. An empirical comparison of seven programming languages. *Computer* 2000;33(10):23-29. DOI: <https://doi.org/10.1109/2.876288>
 39. Souza Júnior EV, Santos GDS, Jesus ALO, et al. Tratamento hemodialítico e seus impactos financeiros no nordeste do Brasil. *Rev Enferm UFPE On Line*. 2019;13. DOI: <http://dx.doi.org/10.5205/1981-8963.2019.239674>
 40. Aly CMC, Reis AT, Carneiro SAM, et al. O Sistema Único de Saúde em série histórica de indicadores: uma perspectiva nacional para ação. *Saúde debate*. 2017;41(113):500-12. DOI: <https://doi.org/10.1590/0103-1104201711312>
 41. Brilhante OMA, Caldas LQ. Gestão e avaliação de risco em saúde ambiental. Rio de Janeiro: Editora FIOCRUZ; 1999.
 42. Moya J, Risi Junior JB, Martinello A, et al. Sala de Situação em Saúde: compartilhando as experiências do Brasil. Brasília, DF: Organização Pan-Americana da Saúde; Ministério da Saúde; 2010.
 43. Rosa L, Mrejen M, Franceschini MC, et al. Instituto de Estudos Para Políticas de Saúde [Internet]. [local desconhecido]: IEPS; 2022 [acesso em 2024 abr 5]. Disponível em: <https://iepsdata.org.br/>
 44. Ministério da Saúde (BR); Conselho Nacional de Saúde. Resolução nº 659, de 26 de julho de 2021. Dispõe sobre a Política Nacional de Informação e Informática em Saúde (PNIIS). *Diário Oficial da União, Brasília, DF*. 2022 jun 15; Edição 113; Seção I:104.
 45. Ministério da Saúde (BR). Estratégia de Saúde Digital para o Brasil 2020-2028. Brasília, DF: Ministério da Saúde; 2020.
 46. Ministério da Saúde (BR), Gabinete do Ministro. Portaria nº 1.434, de 28 de maio de 2020. Institui o Programa Conecte SUS e altera a Portaria de Consolidação nº 1/GM/MS, de 28 de setembro de 2017, para instituir a Rede Nacional de Dados em Saúde e dispor sobre a adoção de padrões de interoperabilidade em saúde. *Diário Oficial da União, Brasília, DF*. 2020 maio 29; Edição 102; Seção I:231.

Recebido em 14/11/2024

Aprovado em 19/06/2025

Conflito de interesses: inexistente

Disponibilidade de dados: os dados de pesquisa estão contidos no próprio manuscrito

Suporte financeiro: não houve

Editor responsável: Alessandro Jatobá