Journal of Health Informatics

J Health Inform. 2026;18:e1824
DOI: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1824

ARTIGO COMPLETO
CBIS 2026

ExtractSUS: metodologia reprodutível para extração automatizada de microdados do Sistema Único de Saúde

ExtractSUS: a reproducible methodology for automated extraction of microdata from the Brazilian Unified Health System

Danyelle Oliveira Fonte[1] ✉, José Alberto Gobbes Cararo[2], Ana Laura de Sene Amâncio Zara[1]

ORCID: 0000-0002-2150-8079 · 0000-0003-1686-4519 · 0000-0001-7012-9078

[1] Universidade Federal de Goiás, Goiânia (GO), Brasil.

[2] Instituto Federal Goiano, Trindade (GO), Brasil.

FICHA DO ARTIGO

TIPO
Artigo completo · CBIS 2026

IDENTIFICADOR
e1824

DOI
10.59681/2175-4411.v18.iEspecial.2026.1824

RECEBIDO
17/05/2026

APROVADO
20/07/2026

PROCESSO DE AVALIAÇÃO
Double Blind Review

EDITOR RESPONSÁVEL
Juliano de Souza Gaspar

ACESSO
Aberto, CC BY 4.0

AUTOR CORRESPONDENTE
Danyelle Oliveira Fonte
danyellefonte@gmail.com

LICENÇA

Este é um artigo publicado em acesso aberto sob a licença Creative Commons CC BY 4.0, que permite uso, distribuição e reprodução em qualquer meio, desde que o trabalho original seja corretamente citado.

CC BY 4.0

RESUMO

Objetivo: Apresentar o ExtractSUS, metodologia reprodutível desenvolvida para extração automatizada de microdados do DataSUS em linguagem Python. **Método:** Trata-se de Design Science Research Methodology que consiste em seis etapas, da identificação do problema à comunicação. A ferramenta opera no Google Colab, baixando arquivos .dbc via FTP do DataSUS e convertendo-os para .parquet e .csv, com parametrização por sistema, Unidades da Federação e período. Comparou-se ExtractSUS, TabNet, TabWin e Microdatasus, e cronometraram-se tempos de extração em cinco sistemas de Goiás (2022). **Resultados:** O ExtractSUS realizou exportação em múltiplos formatos, superando TabWin e TabNet em automação e reprodutibilidade, e foi mais eficiente em bases pequenas; o TabWin foi mais rápido em bases maiores, apesar de possuir mais etapas manuais. **Conclusão:** O ExtractSUS reduz etapas manuais, amplia padronização e reprodutibilidade na extração de dados do SUS, alinhando-se à ciência aberta e à gestão eficiente e transparente de dados públicos.

Descritores: Sistemas de Informação em Saúde; Ciência de Dados; Saúde Digital.

ABSTRACT

Objective: To present ExtractSUS, a reproducible methodology developed for the automated extraction of DataSUS microdata using the Python language. **Method:** This is a Design Science Research Methodology comprising six stages, from problem identification to communication. The tool operates on Google Colab, downloading .dbc files via DataSUS FTP and converting them to .parquet and .csv, with parameterization by information system, federative units, and time period. ExtractSUS, TabNet, TabWin, and Microdatasus were compared, and extraction times were measured across five information systems in Goiás (2022). **Results:** ExtractSUS enabled data export in multiple output formats, surpassing TabWin and TabNet in terms of automation and reproducibility. It also demonstrated greater efficiency when processing smaller datasets, whereas TabWin achieved faster performance with larger datasets despite requiring more manual steps. **Conclusion:** ExtractSUS reduces manual steps, enhances standardization and reproducibility in SUS data extraction, aligning with open science principles and efficient, transparent public data management.

Keywords: Health Information Systems; Data Science; Digital Health.

CBIS 2026 XXI CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAÚDE

Artigo apresentado no XXI Congresso Brasileiro de Informática em Saúde (CBIS 2026), realizado de 23 a 25 de setembro de 2026, em Brasília (DF), e publicado na seção CBIS 2026 do Journal of Health Informatics.

COMO CITAR ESTE ARTIGO: Fonte DO, Cararo JAG, Zara ALSA. ExtractSUS: metodologia reprodutível para extração automatizada de microdados do Sistema Único de Saúde. J Health Inform 2026;18:e1824. doi: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1824

1 INTRODUÇÃO

O Departamento de Informação e Informática do Sistema Único de Saúde (DataSUS) consolidou-se como uma das principais fontes de dados de abrangência nacional no Brasil, reunindo informações estratégicas para a gestão do Sistema Único de Saúde (SUS), a vigilância epidemiológica e a formulação de políticas públicas [1].

Com o intuito de ampliar o acesso às informações, o Ministério da Saúde disponibiliza ferramentas como TabNet e TabWin. O TabWin é um aplicativo que exige instalação local e pode apresentar incompatibilidades com sistemas operacionais atuais, enquanto o TabNet restringe-se a consultas pré-definidas e não oferece acesso aos bancos de dados para análise estatística [2].

O fluxo tradicional de acesso aos microdados do DataSUS, baseado no *download* manual via servidor FTP e posterior conversão por ferramentas auxiliares, apresenta limitações [3]. Embora haja ampla disponibilidade de dados públicos no DataSUS, a utilização desses microdados em pesquisas científicas ainda é limitada devido à reprodutibilidade dos processos de extração e preparação dos dados [4]. Persistem desafios relacionados à padronização, organização e ao tratamento dos dados, o que restringe seu potencial analítico e sua aplicação no apoio à tomada de decisão em saúde [4].

Nos últimos anos, soluções de código aberto têm ampliado as possibilidades de acesso e processamento das bases do DataSUS. Destaca-se o pacote Microdatasus, desenvolvido em linguagem R, que permite o *download* automatizado de arquivos no formato .dbc, a rotulagem de variáveis e o pré-processamento de dados para análises epidemiológicas [5]. Apesar do seu impacto positivo, sua utilização exige conhecimentos em programação, o que representa uma barreira para gestores/as e profissionais da saúde.

Diante desse cenário, evidencia-se uma lacuna metodológica entre a disponibilidade dos dados e sua efetiva utilização. Nesse contexto, o presente estudo tem como objetivo apresentar o ExtractSUS, uma metodologia reprodutível para extração automatizada de microdados do DataSUS, utilizando a biblioteca pySUS versão 1.0.1 [6]. A proposta está alinhada às diretrizes da Estratégia de Saúde Digital para o Brasil 2020–2028, que enfatiza a abertura e o uso qualificado dos dados como pilares para a inovação em saúde [7].

2 MÉTODOS

2.1 Desenho metodológico

O desenvolvimento e a avaliação do ExtractSUS foram conduzidos com base na Metodologia de Pesquisa em Ciência do *Design* (*Design Science Research Methodology*, DSRM) [8]. O percurso metodológico estrutura-se em seis etapas interligadas que guiam o desenvolvimento desde a concepção do método até a disponibilização pública: (i) identificação do problema e motivação; (ii) definição dos objetivos da solução; (iii) projeto e desenvolvimento; (iv) demonstração; (v) avaliação e (vi) comunicação.

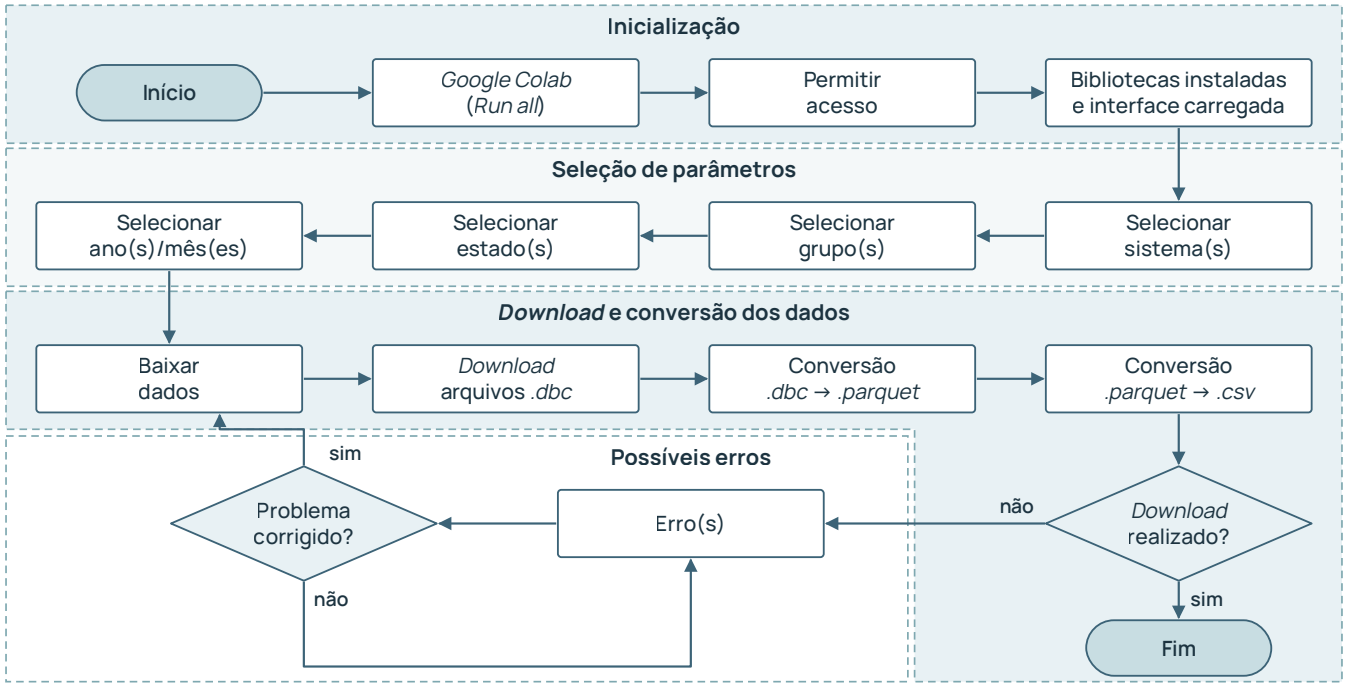
Na primeira etapa**, identificação do problema e motivação**, realiza-se o diagnóstico das barreiras de acesso aos microdados do DataSUS. Identificada a lacuna tecnológica, a segunda etapa trata da definição dos objetivos da solução, na qual se define a meta de desenvolvimento da metodologia reprodutível para acesso e extração automatizada dos microdados do DataSUS. A terceira etapa, **projeto e desenvolvimento**, consiste na arquitetura técnica do ExtractSUS. Opta-se pela linguagem Python e biblioteca pySUS versão 1.0.1 [6], integradas ao ambiente em nuvem Google Colab. Na quarta etapa, **demonstração**, a funcionalidade do ExtractSUS é testada por meio de estudos de casos reais. A quinta etapa, **avaliação**, envolve análise comparativa dos critérios de viabilidade operacional e implicações metodológicas entre o ExtractSUS e os métodos tradicionais (TabNet, TabWin e Microdatasus). Por fim, a sexta etapa, **comunicação**, concretiza-se por meio da disponibilização do método de extração com manual técnico e videoaula explicativa como material suplementar.

2.2 Ambiente computacional

O ambiente computacional adotado é o Google Colab [9], uma plataforma em nuvem que permite a execução de códigos utilizando linguagem de programação Python [10], com recursos computacionais gratuitos, incluindo: (i) acesso à Unidade de Processamento Gráfico (*Graphics Processing Unit*, GPU) com placa T4 NVIDIA, (ii) memória de acesso aleatório (*Random Access Memory*, RAM) entre 12 e 13 GB, (iii) Unidade Central de Processamento (*Central Processing Unit*, CPU) com processador virtual com 2 vCPUs (Intel Xeon ou equivalente) e (iv) armazenamento em nuvem (cerca de 100 GB de armazenamento efêmero). Apesar da disponibilidade da GPU, essa não foi utilizada, uma vez que o processamento dos arquivos é predominantemente em CPU.

A escolha da linguagem Python justifica-se por sua versatilidade e ampla adoção, favorecendo a legibilidade e a produtividade [10]. A biblioteca pySUS v1.0.1 é responsável pelo acesso aos microdados disponibilizados publicamente pelo DataSUS por meio de repositórios FTP, sendo compatível com ambientes Linux [6].

Figura 1. Fluxograma da metodologia de extração automatizada de dados do DataSUS.



Fonte: Elaborado pelos autores (2026).

A extração dos dados é realizada a partir dos Sistemas de Informação em Saúde (SIS) do DataSUS, incluindo 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 Ambulatoriais do SUS (SIA-SUS) e o Sistema de Informações Hospitalares do SUS (SIH-SUS).

O fluxograma metodológico de extração de dados do ExtractSUS é ilustrado na Figura 1. O processo compreende o *download* de arquivos a partir do FTP público do DataSUS no formato .dbc (padrão oficial), seguido de conversão sequencial para .parquet (otimizado para ecossistema Python) e, subsequentemente, para .csv. A parametrização permite selecionar o SIS, grupo específico, Unidades da Federação (UF) e intervalo temporal, gerando subconjuntos de dados sob medida.

A execução inicia-se pela inicialização do ambiente (botão *"Run all"*), concessão de permissão de acesso e instalação automática das dependências. Em seguida, o usuário configura os parâmetros de extração, desencadeando o *download* dos arquivos .dbc e sua conversão automática para os formatos .parquet e .csv. Ao final, os arquivos gerados são consolidados em pasta compactada, disponível tanto localmente quanto em nuvem.

Erros como a indisponibilidade temporária do FTP, arquivos corrompidos ou mudanças de *layout* entre anos podem ocorrer durante o processo de extração. Essas etapas de tratamento dos erros estão previstas na ferramenta e podem ser consultadas no Manual Técnico e Instrucional (https://doi.org/10.5281/zenodo.18379622).

No caso do ExtractSUS, para a utilização do Google Colab [9] gratuito, o particionamento da extração é realizado com 1 UF e/ou ano por vez.

A análise comparativa dos critérios de viabilidade operacional e implicações metodológicas entre o ExtractSUS e os métodos tradicionais (TabNet, TabWin e Microdatasus) foi verificada em julho de 2026, com base na documentação oficial de cada ferramenta e em testes realizados em ambiente Windows. Foram analisadas as seguintes características: tipo de ferramenta, ambiente, interação com usuário, conhecimento de programação, sistema, instalação local, formatos, editável (planilha), filtros customizáveis, código-fonte público, licença de uso, execução por *script*, *download* direto do banco de dados do DataSUS (.dbc), conversão automatizada do formato .dbc para .csv e interface do usuário (configurações para *download*).

Tabela 1. Comparação entre abordagens de acesso aos dados do DataSUS sob a perspectiva da ciência aberta e reprodutibilidade

Dimensão	Características	TabNet	TabWin	Microdatasus	ExtractSUS
Acesso e operação	Tipo de ferramenta	Web	*Desktop*	Código aberto	Código aberto
	Ambiente	Navegador web	Windows	R/RStudio	Google Colab
	Interação com usuário	GUI	GUI	API	GUI
Requisitos técnicos	Conhecimento de programação	Dispensável	Dispensável	Programação em R	Dispensável
	Sistema operacional	Navegador	Windows	Windows/Linux/Mac	Navegador
	Instalação local	Não	Sim	Sim	Não
Formatos de saída	Formatos	HTML	.dbf, .dbc	data.frame (R)	.parquet, .csv
	Editável (planilha)	Não	Parcial	Sim	Sim
	Filtros customizáveis	Sim	Sim	Sim	Sim
Ciência aberta	Código-fonte público	Não	Não	Sim	Sim
	Licença de uso	Proprietária	Proprietária	GPL-3	GPL-3
	Execução por *script*	Não	Não	Sim	Sim
Viabilidade operacional	*Download* direto do banco de dados do DataSUS (.dbc)	Não	Sim	Sim	Sim
	Conversão automatizada do formato .dbc para .csv	Não	Não	Sim	Sim
	Interface do usuário (configurações para *download*)	Sim	Sim	Não	Sim

Nota: *Graphical User Interface* (GUI): interface gráfica do usuário. *Application Programming Interface* (API). Planilha editável indica que os arquivos obtidos podem ser abertos em editores. Edição parcial indica a necessidade de conversor. *General Public License v3.0* (GPL-3): Licença Pública Geral GNU, versão 3.0.

Fonte: Elaborado pelos autores (2026).

Após a caracterização operacional das ferramentas, buscou-se comparar os efeitos práticos desses fluxos no processo de preparação dos microdados para uso analítico, considerando ações manuais, etapas de conversão e nível de conhecimento técnico no uso do ExtractSUS e TabWin. Para comparação do tempo de obtenção dos arquivos, o fluxo manual no TabWin compreende quatro etapas de intervenção direta do usuário: (i) acesso ao servidor FTP/HTTP do DataSUS; (ii) seleção e *download* dos arquivos .dbc; (iii) descompressão individual para .dbf; e (iv) importação e conversão. O tempo foi cronometrado do início do acesso ao servidor até a disponibilização do arquivo em formato analítico (.csv). No TabWin foram consideradas as etapas manuais de seleção dos parâmetros, transferência dos arquivos no site do DataSUS (.dbc) e suas conversões (.dbc → .dbf) e (.dbf → .csv). No ExtractSUS foram consideradas as etapas de seleção dos parâmetros e de execução automática do *script* para *upload*, conversões (.dbc → .parquet → .csv) e *download* dos arquivos (.parquet e .csv). O teste foi realizado em julho de 2026, utilizando um computador Intel Core i7 (2,40 GHz), 16 GB de RAM, 64 bits, Windows 11 Pro; conexão fibra óptica 700 Mbps; Google Colab com Python 3.12.13, pySUS 1.0.1 e numpy 1.26.4; e TabWin 4.1.5.

Este estudo utilizou exclusivamente dados públicos e anonimizados do DataSUS e, portanto, não necessitou de aprovação do Comitê de Ética em Pesquisa.

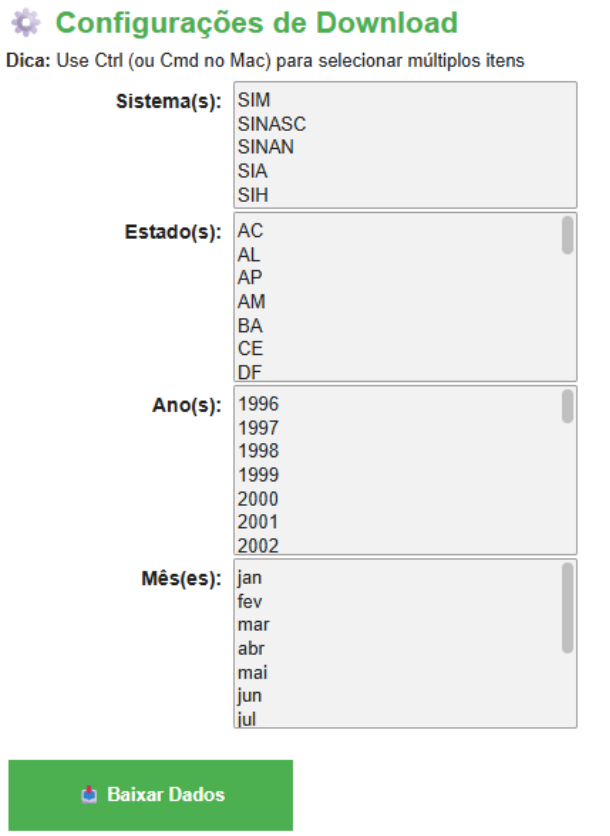
3 RESULTADOS

O ExtractSUS permite a execução de um fluxo padronizado e reprodutível de acesso e extração dos microdados do DataSUS, por meio de linguagem Python. A interface gráfica para configurações de *download* dos arquivos dos SIS é apresentada após a execução no Google Colab. A interface do usuário é composta por células sequenciais nas quais: (1) definem-se os parâmetros de extração (SIS/subgrupos, UFs, anos e meses); (2) pressiona-se o botão "Baixar Dados" para executar o fluxo automatizado que realiza o *download*, conversão e exportação dos dados; e (3) visualiza-se, ao final, um sumário com o número de registros processados e quantidade de arquivos gerados (Figura 2).

Para avaliar qualitativamente a contribuição do ExtractSUS frente às abordagens existentes de acesso aos dados do DataSUS, na Tabela 1 é apresentada uma análise comparativa sistemática entre o ExtractSUS e os principais métodos utilizados: TabNet, TabWin e Microdatasus.

A comparação evidenciou distinção entre as abordagens tradicionais de acesso aos dados do DataSUS e as soluções orientadas à ciência aberta e à reprodutibilidade. Enquanto TabNet e TabWin apresentam limitações como ausência de código-fonte público, licenças proprietárias e reduzida capacidade de customização, as ferramentas Microdatasus e

Figura 2. Interface do usuário no ExtractSUS para configurações de download dos arquivos.



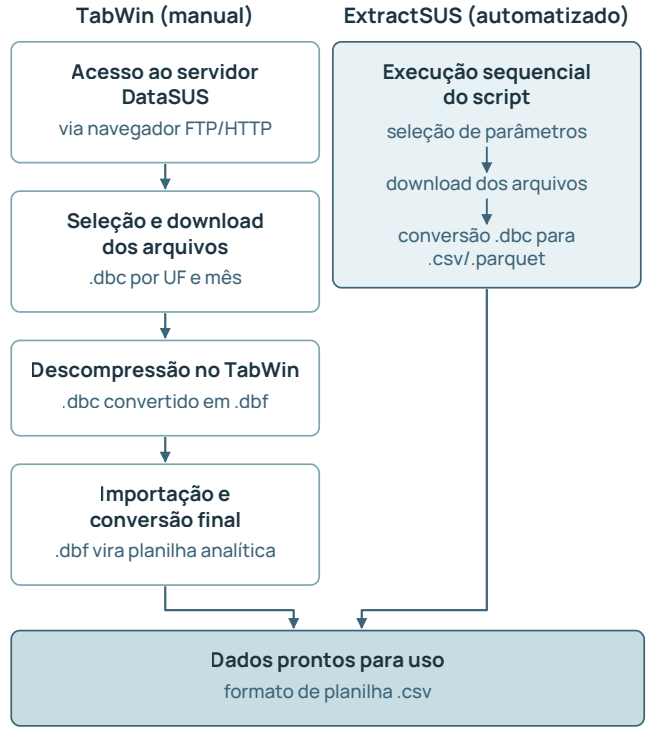
Fonte: Elaborado pelos autores (2026).

ExtractSUS disponibilizam código aberto, suporte à automação via *scripts* e filtros personalizáveis. O Microdatasus destacou-se pela abrangência operacional (Windows, Linux e macOS), integração ao ambiente R/RStudio e acesso por Interface de Programação de Aplicações (API), embora requeira instalação local e conhecimentos de programação. Por sua vez, o ExtractSUS combina características de ciência aberta com maior acessibilidade ao operar em ambiente Google Colab, dispensar instalação local e exportar dados em formato .csv. Em conjunto, esses resultados indicam que as soluções abertas ampliam substancialmente a transparência, a possibilidade de automação e a reprodutibilidade das análises baseadas em dados do DataSUS (Tabela 1).

Na Tabela 1, são apresentadas as características operacionais dos métodos de extração quanto à obtenção de microdados utilizáveis do DataSUS. Embora extrações com TabNet, TabWin e com o pacote Microdatasus permitam diferentes formas de acesso aos dados disponibilizados pelo DataSUS, observam-se diferenças relevantes quanto às características operacionais envolvidas no processo necessário para transformar os arquivos originais em microdados efetivamente utilizáveis para análises estatísticas, com interface do usuário amigável.

Na Figura 3, é apresentado o fluxo de processamento dos dados pelo DataSUS, baseado em múltiplas ações manuais do usuário, e o fluxo automatizado pelo ExtractSUS. Entende-se por ação manual aquela que depende da intervenção direta do usuário e que, portanto, não pode ser reproduzida integralmente via *script*. No TabWin, as quatro etapas manuais identificadas são: (i) acesso ao servidor FTP/HTTP do DataSUS por meio de navegador; (ii) seleção, *download* e conferência visual dos arquivos .dbc de cada UF e competência (mês/ano); (iii) descompressão individual dos arquivos .dbc para o formato .dbf por meio do TabWin; e (iv) importação e conversão final dos arquivos .dbf para formato analítico utilizável em editor de planilhas. No ExtractSUS, essas quatro etapas são executadas de forma sequencial e ininterrupta pelo *script* Python, sem necessidade de intervenção (Figura 3).

Figura 3. Comparação entre os fluxos para obtenção de microdados do DataSUS usando TabWin (manual) e ExtractSUS (automatizado)



Fonte: Elaborado pelos autores (2026).

A fim de comparar a eficiência entre os sistemas, foram cronometrados os tempos de extração de dados no TabWin e no ExtractSUS, apresentados na Tabela 2. O tempo de *download* dos arquivos tende a ser menor no TabWin do que no ExtractSUS. Contudo, em bases com poucos registros, o maior número de etapas manuais exigidas pelo TabWin pode

Tabela 2. Comparação dos tempos globais (em segundos) de extração de dados entre os sistemas TabWin e ExtractSUS para diferentes Sistemas de Informação em Saúde no estado de Goiás, entre janeiro e dezembro do ano de 2022.

Sistema de Informação em Saúde	Grupo / Agravo	Registros (linhas X colunas)	Tempo global (seg)		Δ*
			ExtractSUS	TabWin	
Sistema de Informações sobre Mortalidade	Declaração de Óbito (DO)	47.271 x 87	45	208	+163
Sistema de Informações sobre Nascidos Vivos	Declaração de Nascidos Vivos (DN)	89.746 x 61	47	187	+140
Sistema de Informação de Agravos de Notificação	Meningite (MENI)	20.636 x 131	39	173	+134
Sistema de Informações Ambulatoriais do SUS	Autorização de Procedimento de Alta Complexidade (APAC) de Medicamentos (AM)	726.094 x 51	646	333	-313
Sistema de Informações Hospitalares do SUS	Autorização de Internação Hospitalar (AIH) Reduzida (RD)	369.948 x 113	474	294	-180

Nota: Δ*: tempo_TabWin – tempo_ExtractSUS.

Fonte: Elaborado pelos autores (2026).

tornar o processo global mais demorado, fazendo com que o ExtractSUS apresente melhor desempenho quando se considera o tempo total de extração. Em contrapartida, para arquivos de maior volume, o TabWin tende a ser mais rápido, mesmo diante de sua maior demanda operacional.

Assim, a eficiência relativa dos sistemas varia conforme o tamanho da base analisada e deve ser avaliada considerando todo o fluxo de trabalho, e não apenas o tempo de *download*. Ressalta-se, ainda, que uma parcela substancial do tempo de extração no ExtractSUS decorre do *download* de arquivos no formato .parquet, etapa inexistente no TabWin. Portanto, a comparação direta entre os tempos de extração dos dois sistemas não apresenta completa equivalência operacional.

4 DISCUSSÃO

A proposta busca ampliar, padronizar e democratizar o acesso aos dados, destacando-se por: (i) permitir a extração automatizada sem exigir experiência em programação; (ii) disponibilizar *scripts* de código aberto, reutilizáveis e adaptáveis; e (iii) oferecer abordagem intuitiva, reduzindo erros manuais e aumentando a reprodutibilidade das análises. Nesse sentido, o ExtractSUS posiciona-se como solução intermediária entre a simplicidade das ferramentas oficiais e a robustez de soluções automatizadas, equilibrando acessibilidade e flexibilidade.

A principal contribuição do ExtractSUS reside na automação do *download* de arquivos do DataSUS e na proposta metodológica que padroniza, documenta e torna reprodutível todo o processo de obtenção de microdados de bases públicas. Ao estruturar e documentar todo o fluxo de obtenção dos dados, desde o *download* dos arquivos brutos até sua disponibilização em diferentes formatos, o ExtractSUS estabelece um modelo replicável que pode ser utilizado por pesquisadores, gestores e instituições.

A comparação entre o uso do TabWin e do ExtractSUS evidencia que as diferenças entre fluxos de extração de dados não se limitam à conveniência operacional, mas possuem impacto direto na confiabilidade, na padronização e na reprodutibilidade do uso dos microdados do DataSUS em pesquisas em saúde pública, aspectos alinhados aos princípios da ciência aberta e da reprodutibilidade computacional. Dessa forma, a principal vantagem do ExtractSUS não se relaciona apenas à redução do tempo de *download*, mas à redução das etapas manuais de conversão, à disponibilização imediata dos dados em diferentes formatos (.parquet e .csv) e à padronização do fluxo de extração, fatores que reduzem erros operacionais e barreiras técnicas até a criação do banco de dados para análise.

Um aspecto relevante observado durante a análise refere-se à diferença entre o *download* dos arquivos disponibilizados pelo DataSUS e a efetiva obtenção de microdados em formato utilizável para análise estatística. No TabWin, os arquivos são disponibilizados no formato original *.dbc*, que não é diretamente compatível com a maioria dos ambientes modernos, exigindo etapas adicionais de leitura, conversão e organização realizadas manualmente para transformação em formatos abertos como *.csv* ou estruturados e otimizados para análise como *.parquet* [2]. Essas etapas, embora frequentemente não descritas, demandam conhecimento técnico específico e introduzem variabilidade e uma série de ações externas no processo de preparação dos dados.

Todos os métodos analisados no presente estudo são gratuitos em suas configurações básicas; contudo, o custo-benefício do ExtractSUS depende da escala de processamento. Para extrações pontuais, realizadas por UF e competência (mês/ano), o Google Colab gratuito oferece recursos suficientes, dispensando infraestrutura local e custos adicionais. Em demandas institucionais ou de grande volume, que envolvam múltiplas UFs e períodos, os planos Colab Pro/Pro+ podem ampliar a capacidade computacional, mas

introduzem custos recorrentes. Nesses cenários, o TabWin pode apresentar melhor custo-benefício quando houver computadores compatíveis disponíveis, embora preserve as limitações operacionais e de reprodutibilidade associadas ao fluxo manual. O ExtractSUS pode ser utilizado em qualquer sistema operacional por meio do Google Colab, sendo possível utilizar ferramenta computacional robusta em nuvem, mesmo utilizando computadores com baixos recursos computacionais. A utilização desse ambiente possibilita a padronização da execução dos *scripts*, favorecendo a reprodutibilidade do processo [9].

O ExtractSUS apresenta limitações como: (i) dependência tecnológica da biblioteca pySUS e do Google Colab; (ii) escalabilidade para grandes volumes de dados no plano gratuito do Google Colab; (iii) privacidade, pois embora o DataSUS disponibilize dados anonimizados, a execução em infraestrutura de terceiros pode ser inadequada para instituições com políticas rígidas de governança de dados; (iv) limites de memória RAM no plano gratuito perante a extração simultânea de múltiplos anos em SIS com grandes volumes de registros; e (v) possibilidade de desconexão por inatividade no plano gratuito após 12 horas. No plano Pro+, o tempo máximo é de até 24 horas em execução contínua em segundo plano. Para evitar a desconexão, recomenda-se particionar a extração para SIS com muitos registros. Para mitigar essas limitações, pretende-se, como próxima etapa do projeto, expandir o ExtractSUS para versão *web* desvinculada do Google Colab, proporcionando independência tecnológica.

A padronização do processo de obtenção dos microdados possui implicações diretas para pesquisas que dependem da replicação fiel dos procedimentos de extração dos dados. A ferramenta ExtractSUS surge como resposta a essas limitações. Essa iniciativa está alinhada às boas práticas em computação científica, que enfatizam a necessidade de transparência, reprodutibilidade e rastreabilidade nos processos analíticos. Ao reduzir etapas manuais e suscetíveis a erro, a ferramenta contribui para a melhoria da qualidade metodológica das pesquisas com dados secundários do SUS.

Além disso, a ampliação do acesso e uso de dados em saúde é respaldada pela Lei de Acesso à Informação (Lei nº 12.527/2011), que fortalece a transparência e a democratização das informações públicas [11]. Nesse sentido, ferramentas como o ExtractSUS potencializam a efetivação desse direito ao tornar os dados mais acessíveis e utilizáveis por pesquisadores e gestores. Por outro lado, o uso de dados em saúde deve observar rigorosamente os princípios estabelecidos pela Lei Geral de Proteção de Dados (Lei nº 13.709/2018), especialmente no que se refere ao tratamento de dados sensíveis [12]. Embora o DataSUS disponibilize majoritariamente dados agregados e anonimizados, a utilização dessas informações exige responsabilidade ética e metodológica.

O ExtractSUS está condicionado à disponibilidade e ao escopo dos arquivos públicos, não contemplando sistemas de informação em saúde cujo acesso seja restrito por critérios legais, éticos ou institucionais. De forma semelhante, outros grupos e sistemas que não disponibilizam microdados individuais, como bases exclusivamente tabulares acessadas via TabNet, painéis interativos ou indicadores previamente agregados, também não são passíveis de extração.

5 CONCLUSÃO

O ExtractSUS demonstra que a aplicação de uma metodologia estruturada e orientada ao desenvolvimento de soluções tecnológicas pode transformar de maneira expressiva o acesso aos microdados do DataSUS. Ao integrar, em fluxo único automatizado, etapas que antes dependiam de múltiplas intervenções manuais, o método reduz barreiras técnicas, amplia a padronização e fortalece a reprodutibilidade computacional. A iniciativa se sustenta em um sólido arcabouço legal que garante tanto o direito à saúde quanto o acesso à informação, ao mesmo tempo em que impõe diretrizes para a proteção de dados. Nesse sentido, o uso do ExtractSUS materializa princípios estabelecidos pela Constituição Federal, pela legislação do SUS e pelas normativas de transparência e privacidade. Por fim, o ExtractSUS reforça o papel das tecnologias digitais como instrumentos estratégicos para a consolidação de um SUS orientado por dados, transparente e eficiente.

Contribuição dos autores. Os autores contribuíram igualmente em todas as etapas de construção do artigo: concepção e delineamento do estudo; coleta, análise e interpretação dos dados; redação e revisão crítica do manuscrito; e aprovação da versão final.

Agradecimentos. Os autores agradecem ao Comitê Científico do Congresso Brasileiro de Informática em Saúde (CBIS), ao Programa de Pós-graduação em Assistência e Avaliação em Saúde (PPGAAS) da Universidade Federal de Goiás (UFG) e à Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG), processo SEI nº 2025102670006660, pelo apoio financeiro à realização desta pesquisa.

Financiamento. Os autores não declararam fonte de financiamento.

Disponibilidade de dados e materiais. O código-fonte do ExtractSUS encontra-se disponível publicamente em repositórios no *GitHub* e *Code Ocean*. O manual técnico instrucional e a videoaula explicativa estão depositados no Zenodo. **Repositório do projeto** Google Colab (acesso direto): https://colab.research.google.com/github/zehgobbes/ExtractSUS/blob/main/ExtractSUS_v1.0.ipynb GitHub (versionamento e compartilhamento): https://github.com/zehgobbes/ExtractSUS *Code Ocean* (reprodutibilidade científica): https://codeocean.com/capsule/5884867/tree/v1 https://doi.org/10.24433/CO.5093709.v1 Zenodo (Videoaula): https://doi.org/10.5281/zenodo.18378368 Zenodo (Manual Técnico e Instrucional): https://doi.org/10.5281/zenodo.18379622

Conflitos de interesse. Os autores declaram não haver conflito de interesses.

Declaração de uso de IA. Os autores declaram a utilização de ferramentas de Inteligência Artificial para a revisão gramatical do texto, tradução do resumo para o inglês e produção da Figura 3 (manus.im). A ferramenta chatgpt.com foi utilizada para a correção de erros no código Python.

REFERÊNCIAS

1. Areco KN, Konstantyner T, Bandiera-Paiva P, Balda RCX, Costa-Nobre DT, Sanudo A, et al. Operational Challenges in the Use of Structured Secondary Data for Health Research. Front Public Health. 2021;9:642163. doi: https://doi.org/10.3389/fpubh.2021.642163
2. Silva NP. A utilização dos programas TABWIN e TABNET como ferramentas de apoio à disseminação das informações em saúde. Rio de Janeiro: Fundação Oswaldo Cruz; 2009.
3. Yano KM, Prado SI, Couto W, Zucchi P, Novaes MAP. Limitações no uso do DATASUS como fonte de dados de pesquisas científicas. In: Anais do I Congresso Brasileiro de Saúde Pública On-line: Uma abordagem Multiprofissional. [s.l.]: Revista Multidisciplinar em Saúde; 2021. p. 27. doi: https://doi.org/10.51161/rems/2780
4. Lima LF, Meireles F, Peres L. Tecnologias De Engenharia de Software para o Desenvolvimento de Sistemas de Saúde de Código Aberto: um Mapeamento Sistemático da Literatura. J Health Inform. 2023;15(Especial). doi: https://doi.org/10.59681/2175-4411.v15.iespecial.2023.1079
5. Saldanha RF, Bastos RR, Barcellos C. [Microdatasus: a package for downloading and preprocessing microdata from Brazilian Health Informatics Department (DATASUS)]. Cad Saude Publica. 2019;35(9):e00032419. doi: https://doi.org/10.1590/0102-311X00032419
6. Coelho FC, Baron BC, Fonseca GMC, Reck P, Palumbo D. AlertaDengue/PySUS: Vaccine. Zenodo; 2021. DOI: https://doi.org/10.5281/zenodo.4883502. Documentação do PySUS. 2025. Disponível em: https://pysus.readthedocs.io/en/latest/. Acesso em: 4 nov. 2025.
7. Brasil. Ministério da Saúde. Estratégia de Saúde Digital para o Brasil 2020-2028 [Internet]. Brasília: Ministério da Saúde; 2020 [citado 2026 out 18]. Disponível em: https://bvsms.saude.gov.br/bvs/publicacoes/estrategia_saude_digital_Brasil.pdf
8. Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S. A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems. 2007;24(3):45-77. doi: https://doi.org/10.2753/mis0742-1222240302
9. Google. Google Colaboratory. Disponível em: https://colab.research.google.com/. Acesso em: 27 jul. 2026.
10. Python Software Foundation. Python documentation [Internet]. Wilmington (DE): Python Software Foundation; 2023. Disponível em: https://docs.python.org/3/
11. Brasil. Lei nº 12.527, de 18 de novembro de 2011. Regula o acesso a informações [Internet]. Diário Oficial da União. 2011 nov 18 [citado 2026 out 18]. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm
12. Brasil. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD) [Internet]. Diário Oficial da União. 2018 ago 15 [citado 2026 out 18]. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

NOTAS EDITORIAIS

O Journal of Health Informatics é uma publicação científica mantida pela Sociedade Brasileira de Informática em Saúde, associação científica sem fins lucrativos dedicada ao desenvolvimento da Informática em Saúde e da Saúde Digital no Brasil. O processo editorial é conduzido com independência científica, sem interferência institucional, comercial ou de patrocinadores, e segue as boas práticas internacionais de ética em publicação científica.

Este artigo foi avaliado por pares, em sistema duplo-cego, por especialistas da área. A equipe editorial usou ferramentas de inteligência artificial, com supervisão humana, na triagem inicial, na conferência do texto, das referências e dos números e, quando necessário, no redesenho de figuras, quadros e tabelas, aprovado pelos autores na prova. As ferramentas e as salvaguardas estão na Política de uso de inteligência artificial da revista. As decisões editoriais foram tomadas por pessoas.

SBIS Sociedade Brasileira de Informática em Saúde