ASKLEPION
INFORMAÇÃO EM SAÚDE

ARTIGO

DOI: https://doi.org/10.21728/ asklepion.2026v5n1e-136

Data de submissão: 29/05/2026 Data de aprovação: 28/06/2026 Data de publicação: 29/06/2026

QUALIDADE DE DADOS NO SISTEMA DE INFORMAÇÃO SOBRE MORTALIDADE (SIM) NA PERSPECTIVA DA ANÁLISE EXPLORATÓRIA DE DADOS

Ivan Luiz Marques Ricarte[1]
Universidade Estadual de Campinas
ricarte@unicamp.br

Maria Cristiane Barbosa Galvão[2]
Universidade de São Paulo
mgalvao@usp.br

Resumo

O Sistema de Informação sobre Mortalidade (SIM) é importante instrumento para a definição de estratégias de saúde e definição de políticas públicas, mas para que possa ser efetivamente utilizado seus dados devem oferecer confiabilidade e consistência. Neste estudo, foi aplicada a abordagem de análise exploratória de dados ao conjunto de dados do SIM de 2014, 2019 e 2024. O método de análise exploratória investiga as características de um conjunto de dados sem definição prévia de hipóteses, permitindo observar ocorrência de padrões e consistência dos dados. Embora os dados do SIM analisados apresentem evolução na taxa de preenchimento, com bons valores na maior parte das variáveis, ainda há problemas como campos não preenchidos, inconsistências no preenchimento de campos dependentes entre si, problemas de interoperabilidade com outras fontes de dados, no preenchimento de datas e em aspectos relacionados a determinantes sociais da saúde. O estudo aponta aspectos que podem ser reforçados na implementação da versão eletrônica da Declaração de Óbito, documento fonte para o SIM, mas também mostra que o investimento em capacitação profissional, revisão de instrumentos documentais, harmonização terminológica e fortalecimento das estratégias de interoperabilidade e governança de dados em saúde digital são essenciais.

Palavras-chave: sistema de informação; mortalidade; análise de dados; qualidade da informação; Brasil.

DATA QUALITY IN THE MORTALITY INFORMATION SYSTEM (SIM) FROM THE PERSPECTIVE OF EXPLORATORY DATA ANALYSIS

Abstract

The Mortality Information System (SIM) is an important tool for defining health strategies and public policies, but for it to be effectively used, its data must offer reliability and consistency. In this study, an exploratory data analysis approach was applied to the SIM dataset from 2014, 2019, and 2024. The exploratory analysis method investigates the characteristics of a dataset without pre-defined hypotheses, allowing the observation of patterns and data consistency. Although the analyzed SIM data show improvement in the completion rate, with good values in most variables, there are still problems such as unfilled fields, inconsistencies in the completion of interdependent fields, interoperability problems with other data sources, date completion issues, and aspects related to social determinants of health. The study points to aspects that can be strengthened in the implementation of the electronic version of the Death Certificate, the source document for the Mortality Information System (SIM), but also shows that investment in professional training, review of documentary instruments, terminological harmonization, and strengthening strategies for interoperability and data governance in digital health are essential.

Keywords: information system; mortality; data analysis; information quality; Brazil.

[1] Professor da Universidade Estadual de Campinas (Unicamp), vice-coordenador do Grupo CONFLUÊNCIA, vinculado ao Instituto de Estudos Avançados da USP, com foco em tecnologia e informação em saúde para populações vulneráveis.

[2] Professora da Universidade de São Paulo (USP) e Cientista da Informação. Coordena o Grupo CONFLUÊNCIA. Lidera também o Global Digital Lab: our lab is here! e o CBRIS – Centro Brasileiro de Referência em Informação em Saúde

CALIDAD DE LOS DATOS EN EL SISTEMA DE INFORMACIÓN SOBRE MORTALIDAD (SIM) DESDE LA PERSPECTIVA DEL ANÁLISIS EXPLORATORIO DE DATOS

Resumen

El Sistema de Información sobre Mortalidad (SIM) es una herramienta importante para definir estrategias de salud y políticas públicas, pero para que su uso sea efectivo, sus datos deben ofrecer fiabilidad y consistencia. En este estudio, se aplicó un enfoque de análisis exploratorio de datos al conjunto de datos del SIM de 2014, 2019 y 2024. El método de análisis exploratorio investiga las características de un conjunto de datos sin hipótesis predefinidas, lo que permite observar patrones y consistencia de los datos. Si bien los datos del SIM analizados muestran una mejora en la tasa de completitud, con buenos valores en la mayoría de las variables, aún existen problemas como campos sin completar, inconsistencias en la completitud de campos interdependientes, problemas de interoperabilidad con otras fuentes de datos, problemas de completitud de fechas y aspectos relacionados con los determinantes sociales de la salud. El estudio señala aspectos que pueden fortalecerse en la implementación de la versión electrónica del Certificado de Defunción, el documento fuente del Sistema de Información sobre Mortalidad (SIM), pero también muestra que la inversión en capacitación profesional, la revisión de instrumentos documentales, la armonización terminológica y el fortalecimiento de las estrategias de interoperabilidad y gobernanza de datos en salud digital son esenciales.

Palabras clave: sistemas de información; mortalidad; análisis de datos; calidad de la información; Brasil.

1 INTRODUÇÃO

O Sistema de Informação sobre Mortalidade (SIM) foi um dos primeiros componentes dos Sistemas de Informação do Ministério da Saúde do Brasil, tendo sido desenvolvido em 1975 e informatizado em 1979 (Senna, 2009). Ele tem como fonte de informação os dados obtidos das Declarações de Óbitos (DO), documento distribuído e padronizado nacionalmente. Embora tenha evoluído muito desde sua implantação, tanto em cobertura como em qualidade dos dados, o SIM ainda apresenta deficiências nesses aspectos devido a desigualdades regionais, qualificação de profissionais e infraestrutura dos órgãos municipais de saúde (Rebouças *et al.*, 2025)

Os dados do SIM compreendem informações anonimizadas extraídas da DO de caráter demográfico (como idade, sexo, etnia, município de naturalidade e de residência, nível de escolaridade, situação conjugal, ocupação), sobre a ocorrência do óbito (data e hora, local, município, causas da morte) e informações complementares a depender do tipo de óbito, como dados sobre a mãe se o falecido tem menos de um ano de idade ou sobre as circunstâncias associadas a mortes não naturais (Brasil, 2022).

Ter esses dados disponíveis de forma consistente, ao longo de décadas, como é o caso do SIM, é essencial para a identificação de problemas crônicos de saúde da população e para a definição de políticas públicas visando sanar esses problemas. Para tanto, é preciso que os dados originais e sua codificação no sistema sejam registrados corretamente. Em estudo global realizado há cerca de 20 anos, o SIM foi considerado como um sistema de qualidade intermediária (Mathers *et al.*, 2005) e, desde então, o Ministério da Saúde tem tomado medidas visando à melhoria da qualidade e confiabilidade dos seus dados.

O Ministério da Saúde disponibiliza os dados do SIM desde 1979, tanto por meio da área de transferência de arquivos do Departamento de Informática do Sistema Único de Saúde (Brasil, 2026) como no formato de dados abertos (Portal Brasileiro de Dados Abertos, 2022). Além de permitir a análise e divulgação dos dados em plataformas como a Tabnet, do DATASUS, ou a Plataforma de Ciência de Dados aplicada à Saúde (Fundação Oswaldo Cruz, 2018), isso possibilita que vários estudos sejam realizados não apenas sobre análises de causas de morte por grupos ou regiões geográficas específicas, mas também sobre a qualidade de dados no sistema, como consolidado em revisão recente de Rebouças *et al*. (2025). Nesse trabalho, estudos que contemplaram dimensões de qualidade de dados como acessibilidade, clareza metodológica, cobertura, completitude, confiabilidade, consistência, não-duplicidade, oportunidade, validade e avaliação de causas mal definidas foram analisados e sintetizados.

O presente estudo tem por objetivo explorar os dados do SIM buscando identificar padrões de preenchimento, inconsistências estruturais, dependências lógicas, ambiguidades semânticas e limitações de interoperabilidade presentes nas variáveis armazenadas no sistema, de modo a ampliar a compreensão sobre os desafios relacionados à qualidade da informação em sistemas nacionais de informação em saúde.

2 MÉTODO

Trata-se de uma pesquisa quantitativa, exploratória e documental, baseada em análise secundária de dados abertos do SIM disponibilizados pelo Ministério da Saúde brasileiro por meio do Portal Brasileiro de Dados Abertos (2022). Foram selecionados para análise os conjuntos de dados referentes a três anos, incorporando o período mais recente com dados consolidados disponíveis no momento da coleta (ano de 2024) e dados de cinco (ano de 2019) e dez anos (ano de 2014) antes, para observar se há evolução no preenchimento dos dados.

O estudo adotou a abordagem da Análise Exploratória de Dados (Pearson, 2018) proposta por John Tukey em 1977, compreendida como uma estratégia investigativa orientada à descoberta de padrões, anomalias, inconsistências e relações entre variáveis, sem o emprego inicial de modelos estatísticos inferenciais ou hipóteses confirmatórias previamente definidas. É uma abordagem já reconhecida na área da saúde; a base de dados PubMed registra mais de 1200 trabalhos com o termo "*exploratory data analysis*" desde 1977, com mais de 100 artigos por ano nos últimos anos. Nessa abordagem, a investigação é conduzida de forma iterativa, permitindo que os próprios dados direcionem a identificação de comportamentos inesperados, incoerências lógicas e fragilidades estruturais associadas ao processo de produção e registro da informação.

Para interpretação dos dados, utilizaram-se como referências complementares o modelo oficial da Declaração de Óbito, o Manual de Instruções para o Preenchimento da Declaração de Óbito do Ministério da Saúde (Brasil, 2022) e o Dicionário de Dados do SIM, disponibilizado no Portal Brasileiro de Dados Abertos juntamente com os dados. Inicialmente, realizou-se a inspeção da estrutura geral das bases, contemplando identificação de variáveis, tipos de dados, padrões de codificação e relações condicionais entre campos. Em seguida, foram conduzidos procedimentos de limpeza, transformação, padronização e sumarização dos dados, incluindo tratamento de campos vazios, análise de códigos de preenchimento ignorado, conversão de formatos temporais e inspeção de múltiplos códigos presentes em campos clínicos. As variáveis foram exploradas quanto à completitude, coerência lógica entre campos dependentes,

consistência temporal, adequação às regras de preenchimento previstas nos documentos normativos e compatibilidade com sistemas externos de classificação, incluindo CID-10, CNES, CBO e códigos municipais do IBGE.

A análise exploratória foi conduzida utilizando-se o software R versão 4.6.0, com uso dos pacotes *dplyr* para manipulação de dados, *ggplot2* para visualização gráfica e *lubridate* para tratamento de variáveis temporais (Wickham, 2016).

3 RESULTADOS

O estudo foi realizado em maio de 2026. Nesse momento, os dados do SIM referente ao ano de 2025 ainda não estavam disponibilizados por completo e, por esse motivo, foram analisados os dados de 2024, 2019 e 2014.

O primeiro aspecto observado foi a diferença entre a quantidade de campos na DO e a quantidade de variáveis no SIM. O formulário da DO tem 59 campos, dos quais muitos não são armazenados no SIM para garantir a anonimização de dados pessoais referentes ao óbito. No entanto, os dados do SIM em 2024 continham 88 variáveis e 87 variáveis em 2014 e 2019. Como o formulário da DO tem atualizações periódicas, a última delas realizada em 2011, essas variáveis adicionais devem estar relacionadas a versões antigas, além de alguns campos com dados da aplicação, como origem dos dados e versão do sistema.

Após a seleção e conversão dos campos, usando como referência o manual para preenchimento da DO, foram produzidos para a análise detalhada dois conjuntos de dados com 39 variáveis cada, sintetizadas no Quadro 1 com suas respectivas taxas de preenchimento nos anos analisados.

Quadro 1 - Síntese das variáveis do SIM com base na versão corrente da DO: taxas de preenchimento

Variável	Tipo	2014	2019	2024
(N)	(Quantidade de registros no conjunto de dados)	1.227.039	1.349.801	1.532.015
Tipo de óbito	Categórica (fetal, não fetal)	100%	100%	100%
Data de óbito	Dia/mês/ano	100%	100%	100%
Hora de óbito	Hora:minuto	94,0%	95,8%	96,5%
Naturalidade	Código numérico (Nacionalidade/UF, 3 dígitos)	90,8%	94,5%	95,9%
	Código numérico[1] (Município brasileiro, 6 dígitos)	99,2% (n=1.098.270)	99,7% (n=1.257.268)	99,7% (n=1.454.411)
Data de nascimento	Dia/mês/ano	99,5%	99,7%	99,8%
Idade	Valor numérico (minutos a anos)	100%	100%	100%
Variável	**Tipo**	**2014**	**2019**	**2024**
Sexo	Categórica (masculino, feminino, ignorado)	100%	100%	100%
Raça/cor	Categórica (branca, preta, amarela, parda, indígena)	95,1%	97,4%	98,7%

Situação conjugal	Categórica (solteiro, casado, viúvo, separado judicialmente ou divorciado, união estável, ignorado)	92,3%	94,5%	95,8%
Escolaridade	Nível: Categórica (sem escolaridade, fundamental I, fundamental II, ensino médio, ensino superior incompleto, ensino superior, ignorado)	86,0%	91,8%	93,6%
	Série[2]: Categórica (1 a 8)	25,4% (n=658.003)	39,5% (n=796.954)	46,3% (n=963.347)
Ocupação habitual[3]	Código numérico (6 dígitos)	80,1% (n=1.178.893)	87,0% (n=1.136.322)	89,0% (n=1.494.551)
Município de residência	Código numérico (6 dígitos)	100%	100%	100%
Local de ocorrência do óbito	Categórica (hospital, outro estabelecimento de saúde, domicílio, via pública, outros, aldeia indígena, ignorado)	100%	100%	100%
Estabelecimento de saúde[4]	Código numérico (7 dígitos)	99,97% (n=879.776)	100% (n=991.068)	100,01% (n=1.123.690)
Município de ocorrência do óbito	Código numérico (6 dígitos)	100%	100%	100%
Dados da mãe, se falecido tem menos de um ano:		(n=38.430)	(n=35.293)	(n=30.020)
Idade da mãe[5]	Valor numérico (anos ou 99, se ignorado)	87,0%	90,5%	91,7%
Escolaridade da mãe[5]	Nível: Categórica (como Escolaridade)	85,4%	89,2%	89,7%
	Série: Categórica (como Escolaridade)	46,1% (n=25.271)	55,3% (n=23.615)	59,3% (n=19.782)
Ocupação da mãe[5]	Código numérico (6 dígitos)	72,8%	80,6%	81,6%
Número de filhos tidos vivos[5]	Valor numérico (quantidade ou 99, se ignorado)	85,9%	89,7%	91,2%
Número de filhos nascidos mortos[5]	Valor numérico (quantidade ou 99, se ignorado)	30,3%	37,0%	39,1%
Semanas de gestação[5]	Valor numérico	82,2%	87,7%	89,5%
Tipo de gravidez[5]	Categórica (única, dupla, tripla ou mais, ignorada)	90,8%	93,0%	93,5%
Tipo de parto[5]	Categórica (vaginal, cesáreo, ignorado)	90,4%	92,6%	93,0%
Morte em relação ao parto[5]	Categórica (antes, durante, depois, ignorado)	89,9%	91,8%	92,0%
Peso ao nascer[5]	Valor numérico (gramas)	86,7%	89,3%	90,1%
Situação gestacional[6]	Categórica (na gravidez, no parto, no abortamento, até 42 dias após o parto, de 43 dias a 1 ano após o término da gestação, não ocorreu nesses períodos, ignorado)	123,4% (n=70.579)	81,7% (n=101.508)	87,5% (n=103.128)
Assistência médica	Categórica (sim, não, ignorado)	70,0%	69,4%	69,6%
Diagnóstico confirmado por necropsia	Categórica (sim, não, ignorado)	72,3%	71,0%	71,5%
Causa básica	Código CID	100%	100%	100%
Condição do médico atestante	Categórica (assistente, substituto, IML, SVO, outro)	88,5%	91,3%	93,4%
Variável	**Tipo**	**2014**	**2019**	**2024**
Município do SVO ou IML[7]	Código numérico (6 dígitos)	96,5% (n=250.664)	100,1% (n=249.447)	99,0% (n=276.733)
Data do atestado	Dia/mês/ano	98,4 %	99,3%	99,7%

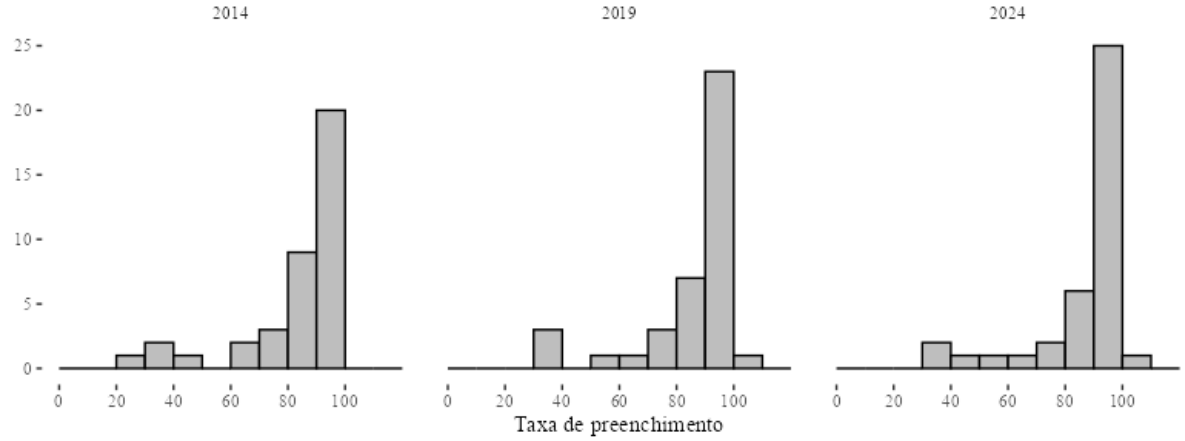
Prováveis circunstâncias[8]	Categórica (acidente, suicídio, homicídio, outros, ignorado)	99,9% (n=156.861)	99,8% (n=142.800)	99,8% (n=159.534)
Acidente de trabalho[9]	Categórica (sim, não, ignorado)	69,3% (n=71.795)	73,7% (n=60.215)	76,1% (n=73.821)
Fonte da informação[8]	Categórica (ocorrência policial, hospital, família, outra, ignorado)	79,3% (n=156.861)	79,3% (n=142.800)	80,6% (n=159.534)
Tipo de ocorrência do óbito[8]	Categórica (via pública, endereço de residência, outro domicílio, estabelecimento comercial, outros, ignorada)	37,4% (n= 156.861)	39,7% (n=142.800)	39,9% (n=159.534)

[1] Preenchido quando o código da nacionalidade é de uma unidade da federação.
[2] Preenchido quando o nível de escolaridade é educação básica (Fundamental 1: 1 a 4; Fundamental II: 5 a 8; Médio: 1 a 3).
[3] Preenchido se o falecido tem pelo menos 5 anos de idade.
[4] Preenchido se local de óbito é hospital ou outro estabelecimento de saúde.
[5] Preenchidos se falecido tem menos de um ano de idade.
[6] Preenchido se falecido é do sexo feminino e em idade fértil.
[7] Preenchido se a condição do médico atestante é Instituto Médico Legal (IML) ou Serviço de Verificação de Óbito (SVO).
[8] Preenchidos se a causa da morte está relacionada ao Capítulo XX da CID-10.
[9] Preenchido se a provável circunstância do óbito é acidente.

Fonte: Elaborado pelos autores (2026).

As taxas de preenchimento para cada variável, nos três anos analisados, são próximas, o que sugere que eventuais padrões de preenchimento são recorrentes e não eventuais. Deve-se observar que esses valores se referem simplesmente às taxas de preenchimento, sem analisar se o preenchimento está consistente ou correto. É possível observar que, a cada ano, mais campos alcançam uma maior taxa de preenchimento, como é possível observar na Figura 1.

Figura 1 - Evolução das taxas de preenchimento



Fonte: Elaborado pelos autores (2026).

Na sequência, são apresentados aspectos relacionados a algumas dessas variáveis que emergiram da análise exploratória. Quando valores são usados para ilustração, referem-se ao ano de 2024.

3.1 TIPO DE ÓBITO

Nos três anos analisados, todos os óbitos registrados no SIM são do tipo “não fetal”. Óbitos fetais são armazenados em outro sistema de informação.

3.2 NATURALIDADE E CÓDIGOS DE MUNICÍPIOS

O campo Naturalidade da DO é mapeado a duas variáveis do SIM, uma com o código do país ou unidade da federação (3 dígitos) e outra com o código do município, caso o falecido seja natural do Brasil.

Não foi possível identificar, na documentação do SIM ou em outras fontes de dados utilizadas em sistemas do SUS, qual é a codificação utilizada para representar outros países que não o Brasil. Quando a nacionalidade é brasileira, um código com o padrão “8XX” é utilizado, sendo os dois últimos dígitos o código usado pelo IBGE para identificação da unidade da federação ou “00”, provavelmente adotado quando essa informação é ignorada.

Para o registro do campo de município é utilizado o código do IBGE; esse código tem sete dígitos, sendo os dois primeiros identificadores da unidade da federação e o sétimo é um dígito verificador. No SIM, apenas os seis primeiros dígitos são registrados (o dígito verificador não é utilizado). Essa codificação é também utilizada para as variáveis Município de residência, Município de ocorrência do óbito e Município do SVO ou IML. Em relação à codificação oficial do IBGE, é utilizada uma extensão para representar um município desconhecido, com os dois primeiros dígitos representando a unidade da federação e os demais dígitos todos zero.

3.3 ESCOLARIDADE

A escolaridade do falecido e a escolaridade da mãe do falecido que tenha até um ano de idade são representadas pela combinação de dois dados, nível (Fundamental I ou II, Médio, Superior) e, no caso da educação básica, qual a última série concluída. Há ainda as opções “sem escolaridade” e “ignorado”.

Apesar de ter a opção “ignorado” (n=139.266, 9,1% para os falecidos; n=1.542, 5,1% para as mães), esse campo é muitas vezes deixado sem nenhum valor (n=97.549, 6,4% para os falecidos; n=3.084, 10,3% para as mães). Outro fator de incerteza em relação a esse dado é que nem sempre o campo com a última série concluída da educação básica é preenchido. No caso da escolaridade do falecido, há 237.948 (15,5%) registros assinalados com Ensino Fundamental I e com a indicação da última série (1 a 4) preenchida contra 240.787 (15,7%) registros com o

mesmo nível, mas sem indicação da série concluída. Os níveis de não preenchimento comparados aos registros preenchidos são ainda mais altos para o Ensino Fundamental II, com 110.621 (7,2%) registros com indicação de série e 132.889 (8,7%) sem indicação, e para o Ensino Médio, com 97.672 (6,4%) com indicação e 143.430 (9,4%) sem indicação de série concluída.

Um problema adicional na escala adotada está na imprecisão associada às mudanças decorrentes de reformas educacionais, como a de 2006 que define o ensino fundamental com nove anos em vez de oito séries. Portanto, há falecidos que fizeram a educação básica com ciclos de onze anos e, entre os mais jovens, de doze anos. A escala utilizada não captura adequadamente essas diferenças.

3.4 OCUPAÇÃO HABITUAL

Este campo recebe um código numérico associado à Classificação Brasileira de Ocupações (CBO) do Ministério de Trabalho e Emprego. No entanto, nem todos os valores usados no contexto do SUS estão presentes nessa classificação, havendo códigos adicionais que precisam ser buscados em outras aplicações do Sistema Único de Saúde (SUS).

Segundo o manual de instruções, esse campo só deve ser preenchido se o falecido tiver pelo menos cinco anos de idade. Outra observação, explicitada também no próprio formulário da DO, é que se o falecido fosse aposentado ou desempregado, a ocupação anterior deveria ser informada.

O primeiro problema observado é o preenchimento indevido. Em 2024, há 58 registros que tem valores nesse campo para falecidos com 4 anos idade ou menos, com ocupações como estudante, dona de casa, desempregado (ou não possível obter ocupação habitual), aposentado/pensionistas e ocupação não identificados. Aparentemente, os declarantes preencheram esse campo com a ocupação da mãe, o que deve ser feito em campo específico e apenas quando o falecido tem menos de um ano de idade, ou adotaram um código genérico para ocupação desconhecida mesmo quando não é necessário preencher o campo.

Outro aspecto a ser observado é que as duas ocupações mais frequentes, compreendendo 39,2% do total, não são ocupações presentes na CBO – aposentado ou pensionista (n=403.787, 26,4%) e dona de casa (n=195.880, 12,8%). O aspecto mais grave em relação a esse dado é que o formulário indica explicitamente que, no caso do falecido ser aposentado, o campo deve ser preenchido com a ocupação anterior.

3.5 ESTABELECIMENTO DE SAÚDE

O campo da DO denominado “Estabelecimento” só deve ser preenchido quando o Local de ocorrência do óbito for “hospital” ou “outro estabelecimento de saúde”, no campo anterior. No entanto, em 2024 há 1.123.690 (73,3%) registros nos quais o local de ocorrência de óbito é assinalado como uma dessas duas opções, mas há 58 registros a mais para os quais um código de estabelecimento foi preenchido. Por esse motivo, no Quadro 1 a taxa de preenchimento para esse campo é superior a 100%. A inspeção dos dados mostra que nesses registros o local de ocorrência do óbito foi apontado como “ignorado” e o código de estabelecimento utilizado foi “9999999”, um código CNES inválido.

3.6 DADOS DA MÃE, GRAVIDEZ E PARTO

O bloco IV da DO só deve ser preenchido quando o falecido tem menos de um ano de idade, o que em 2024 corresponde a 30.020 ocorrências. Há apenas uma ocorrência adicional em 2024 na qual a idade do falecido foi registrada no SIM como “ignorada” (código de idade 999) e os dados desse bloco foram preenchidos. No entanto, há 2.500 ocorrências (8,3%) nos quais esses campos deveriam ser preenchidos e não foram.

3.7 SITUAÇÃO GESTACIONAL

Este é o primeiro campo do Bloco V, apresentado com o título “Óbito de mulher em idade fértil”. Segundo o título e as instruções de preenchimento, este campo só deve ser preenchido se o óbito que está sendo declarado for de uma mulher e, adicionalmente, que ela esteja em idade fértil, ou seja, no período entre a primeira menstruação e a menopausa.

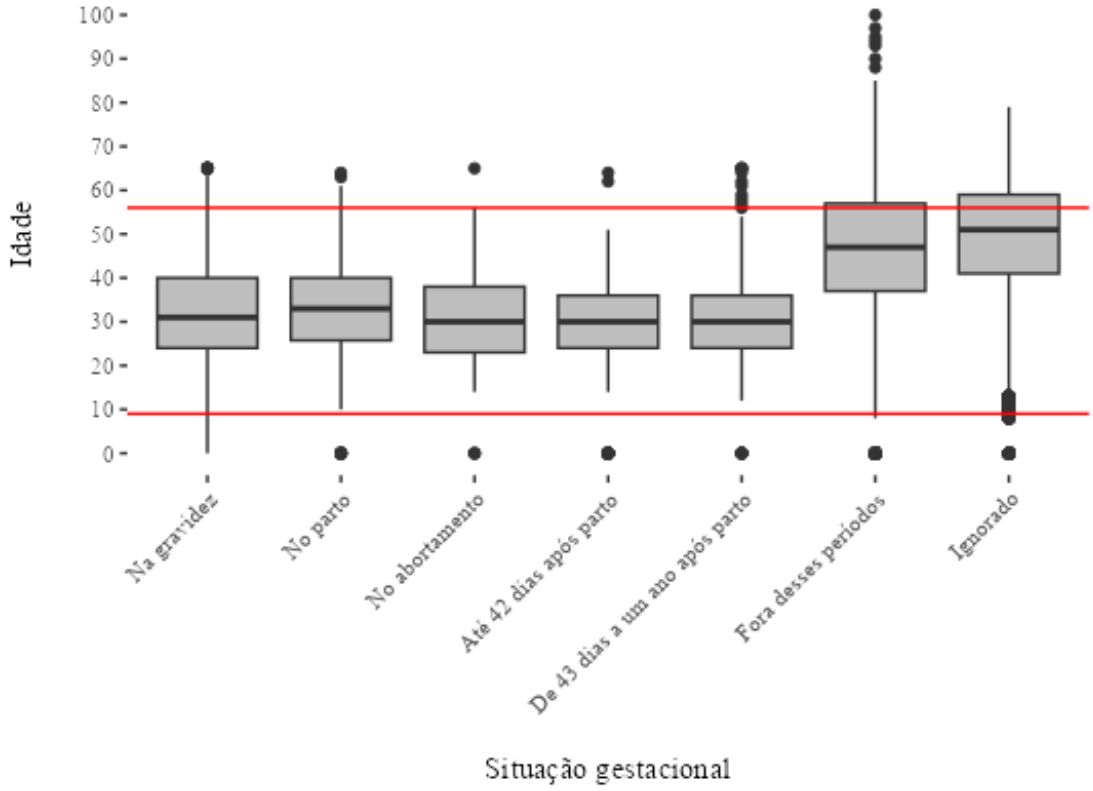
Embora o período correspondente à idade fértil não tenha uma definição precisa, é possível usar como estimativa os dados com as idades das mães de falecidos de até um ano de idade. Em 2024, essa faixa vai dos 9 aos 56 anos.

A primeira inconsistência observada nesse campo é o preenchimento em situações nas quais o sexo atribuído ao falecido não é “Feminino”. Em 2024, isso ocorre em 73 situações, havendo 62 ocorrências onde o sexo é definido como “ignorado” e em 11 ocorrências o falecido é do sexo masculino.

A outra inconsistência é o preenchimento do campo associado a óbitos de mulheres fora da idade fértil. Há 1.460 registros nos quais esse campo está preenchido para óbitos de meninas com menos de um ano de idade, além de 36 registros para mulheres com pelo menos 100 anos

de idade. A Figura 2 mostra como há dados preenchidos nesse campo mesmo para mulheres (potencialmente) fora da idade fértil, cuja estimativa para 2024 está marcada no gráfico com as linhas horizontais vermelhas.

Figura 2 - Distribuição de idade das mulheres para as quais o campo Situação gestacional foi preenchido em 2024.



Fonte: Elaborado pelos autores (2026).

3.8 CAUSAS DA MORTE

O Quadro 1 apresenta uma variável do SIM, Causa básica, que está com preenchimento de 100% e com um único código CID por registro. Não há uma indicação clara no Dicionário de Dados sobre qual é a origem dessa informação a partir da DO.

Na DO, o campo “Causas da morte” oferece seis linhas, com a observação de que apenas um diagnóstico deve ser anotado por linha, com a seguinte lógica de preenchimento: a primeira linha é a causa terminal, com as linhas seguintes estabelecendo uma cadeia de causalidade (“como consequência de”), onde a quarta linha deve estabelecer a causa básica. Há duas linhas adicionais para outros diagnósticos que podem ter contribuído para o óbito sem estar na cadeia direta de causalidade. Há um campo adicional para indicar o tempo aproximado entre o início da doença e a morte (que não é registrado no SIM) e outro para o código CID que, segundo o

manual de instruções, não deve ser preenchido pelo médico declarante. Apenas o código CID é registrado no SIM.

Embora não esteja representado no Quadro 1, as variáveis referentes a essas linhas foram analisadas. O que se observa nos dados do SIM é que a restrição de um diagnóstico por linha não é respeitada. Em 2024 a primeira linha (causa terminal), por exemplo, tem um único código CID em 1.446.213 ocorrências (94,3%), mas não foi preenchida em 46.783 (3,1%) registros; tem dois códigos CID em 35.231 (2,3%) registros; três códigos em 2.549 (0,2%) registros; e quatro códigos em 1.239 (0,1%) registros. Similarmente, a segunda linha não foi preenchida em 355.280 (23,2%) registros, a terceira linha, em 801.926 (52,3%) registros e a quarta linha, que deveria ser a causa básica, em 1.221.847 (79,8%) registros, e em todas essas linhas há registros com múltiplos códigos. Na variável que registra as comorbidades (Parte II) há registros com até nove códigos CID, enquanto o formulário dedica espaço para dois diagnósticos complementares apenas.

3.9 DATA DO ATESTADO

Em 2024, esse campo contém três datas inconsistentes, sendo duas anteriores a 2024 (27/02/2002 e 18/01/2023) e uma em futuro remoto (15/02/2202). A inspeção dos dados mostra que nos dois primeiros casos houve provavelmente um erro no preenchimento do ano apenas, pois as datas de óbitos são respectivamente 27/02/2024 e 18/01/2024. Já a data de óbito para o terceiro caso é 15/03/2024, então não é possível afirmar se houve erro no preenchimento do mês e do ano (03, 2024) ou se, como em outros casos, foi um atestado que demorou a ser emitido e o erro de preenchimento foi só no ano (2025) – o tempo médio de emissão observado em 2024 foi de 49 dias, mas há casos em que o atestado levou mais de 600 dias para ser emitido. Portanto, nem sempre erros de preenchimento podem ser resolvidos num processo de limpeza dos dados.

Outra inconsistência foi observada ao comparar a data de óbito com a data de emissão do atestado: há 95 casos em que a data do atestado precede a data do óbito.

3.10 MORTE POR CAUSAS EXTERNAS

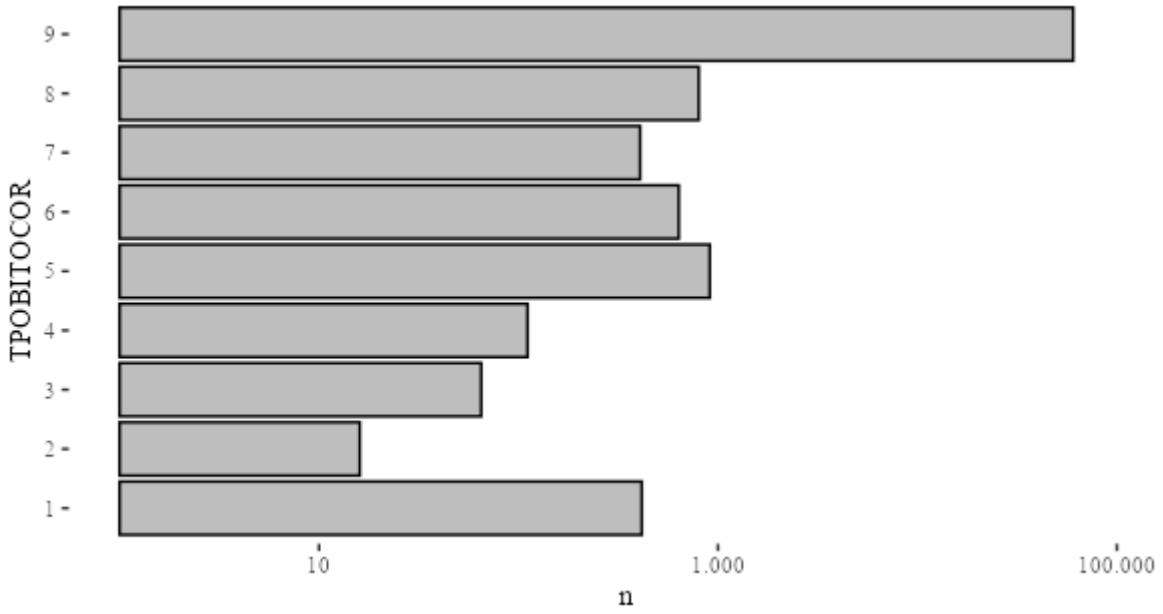
O Bloco VII da DO registra dados sobre as prováveis circunstâncias de morte não natural, cuja causa de morte esteja associada ao Capítulo XX da CID-10 (Causas externas de morbidade e de mortalidade), com códigos entre V01 e Y98.

Considerando apenas os registros nos quais a causa básica de morte está nessa faixa de códigos CID (n=159.534), a taxa de preenchimento das prováveis circunstâncias é alta (99,8%).

No entanto, quando a circunstância é um acidente (o que ocorre em 73.821 casos), é importante, para fins legais e para assegurar direitos de familiares do falecido, a indicação se o acidente foi um acidente de trabalho. No entanto, há muitos registros de mortes por acidente sem essa informação – a taxa de preenchimento é de 76,1%. Observou-se ainda o preenchimento inadequado do campo referente a acidente de trabalho quando a provável circunstância não foi um acidente – no caso, três suicídios, um marcado como "não" (acidente de trabalho) e dois como "ignorado".

No entanto, o caso mais grave no SIM em relação a esse bloco da DO está associado ao campo "Tipo de local de ocorrência do acidente ou violência". Segundo o formulário da DO e o Dicionário de Dados do SIM, essa variável (denominada TPOBITOCOR no SIM) pode assumir valores de 1 a 5, além do valor 9 (ignorado). A inspeção dos dados, no entanto, mostra que essa variável assume valores de 1 a 9, como mostra a Figura 3. O mesmo padrão de ocorrência de valores inconsistentes com a definição foi observado em 2014 e em 2019.

Figura 3 - Distribuição de valores para a variável TPOBITOCOR em 2024



Fonte: Elaborado pelos autores (2026)

4 DISCUSSÃO

O SIM é um dos marcos dentro dos sistemas de informação sobre saúde do Ministério da Saúde brasileiro. Os seus dados constituem importante fonte de informação para pesquisas e para a definição de políticas públicas, motivo pelo qual é importante que seus dados sejam confiáveis e de qualidade exemplar.

O SIM apresenta um conjunto de dados anonimizados obtidos a partir das declarações de óbitos emitidas em todo o território nacional. Há mais de 80 variáveis armazenadas no SIM, com 42 dessas variáveis correspondendo a 39 campos presentes no modelo atual da declaração de óbito que não contêm informação que permitam identificar os indivíduos envolvidos (falecidos, pais, médicos) ou seus endereços precisos.

A análise exploratória dos dados armazenados no SIM mostrou que a taxa de preenchimento para esses campos é alta e com ampla cobertura territorial, mostrando a consolidação do modelo unificado de declaração de óbito adotado há cerca de 50 anos. Considerando as dimensões continentais do Brasil e sua diversidade regional, esse simples fato mostra a relevância das iniciativas relacionadas aos sistemas de informação em saúde de âmbito nacional.

Por outro lado, o estudo mostrou que há padrões recorrentes de inconsistências associados tanto ao preenchimento das Declarações de Óbito quanto às próprias limitações presentes nos mecanismos de representação computacional e codificação adotados pelo sistema. Na sequência, esses padrões são detalhados de acordo com seus eixos analíticos.

4.1 PADRÕES DE COMPLETITUDE E AUSÊNCIA DE DADOS

Embora as instruções de preenchimento da DO explicitem que nenhum campo deve ser deixado em branco, o mesmo cuidado não é aplicado às variáveis do SIM. Algumas das variáveis apresentam apenas preenchimento parcial, com taxas de preenchimento inferior a 40%, como o número de filhos nascidos mortos (para mães de falecidos com menos de um ano de idade), o tipo de ocorrência de óbito (para mortes por causas não naturais) e a última série concluída (para escolaridade de nível fundamental ou médio).

Não há uma diretriz clara sobre a codificação de campos não preenchidos. Em alguns casos, há códigos para especificar que um valor é ignorado e, mesmo quando o valor deveria ser preenchido por condições anteriores do formulário, o campo é deixado em branco ou não é preenchido no SIM.

4.2 INCONSISTÊNCIAS LÓGICO-CONDICIONAIS

Há algumas situações nas quais os valores atribuídos a variáveis são incompatíveis com a idade. Um exemplo dessa situação ocorre na variável Ocupação habitual, que deve ser preenchida apenas se o falecido tem cinco anos ou mais de idade, mas há valores preenchidos para falecidos de quatro anos ou menos. Outro exemplo ocorre na variável Situação gestacional,

que deve ser preenchida apenas quando o óbito é de uma mulher em idade fértil, mas há valores atribuídos a essa variável para mulheres de zero a cem anos de idade.

Nessa mesma variável de Situação gestacional há inconsistências relacionadas ao gênero do falecido. Embora a variável só deva ser preenchida quando o óbito é de mulheres em idade fértil, há registros de preenchimento para falecidos com sexo masculino ou ignorado.

Há situações em que a variável Estabelecimento de saúde tem um valor atribuído embora o óbito não tenha ocorrido em estabelecimento de saúde.

4.3 INCONSISTÊNCIAS SEMÂNTICAS E CLASSIFICATÓRIAS

Os campos referentes a Causas da morte (códigos CID) não estão consistentes com as instruções de preenchimento da declaração de óbito, que indicam que cada linha deve ter um único diagnóstico. Há milhares de registros com mais de um código por linha, chegando a até quatro códigos nas linhas da Parte I e a até nove códigos na Parte II, embora o formulário reserve apenas duas linhas para essa parte. Isso sugere que, ao contrário das recomendações, os médicos declarantes estão preenchendo o campo onde deveria estar a descrição da causa da morte diretamente com códigos da CID. Há apenas uma variável (Causa básica) que tem apenas um código CID, mas não está clara qual é a origem dessa informação, uma vez que a causa básica da morte deve ser especificada na linha d da Parte I, mas há divergências entre os valores atribuídos a essas duas variáveis.

Os campos referentes à escolaridade (nível e série concluída) também são passíveis de revisão. Desde a promulgação da Lei 11.274/2006, o ensino fundamental passou de oito séries para nove anos, o que demandaria uma adequação no formulário para indicar de forma mais precisa a escolaridade do falecido ou de sua mãe, quando for o caso.

Com relação à variável Ocupação habitual do falecido, as instruções de preenchimento da DO explicitam que caso o falecido seja aposentado o registro deve indicar a profissão anterior. No entanto, "Aposentado" é a ocupação mais frequente nessa variável.

4.4 PROBLEMAS DE INTEROPERABILIDADE

Apesar do Dicionário de Dados indicar que os campos com código do município (de naturalidade, de óbito, de ocorrência) são de sete dígitos, todos esses dados têm efetivamente seis dígitos. O sétimo dígito é um dígito verificador e, para que ocorra a integração com os dados do SIM, precisa ser desconsiderado das tabelas obtidas no IBGE. Adicionalmente, os códigos adotados para representar uma unidade da federação com município desconhecido não

fazem parte dessas tabelas e precisam ser adicionados manualmente para poder ser incluídos em futuras análises. Adicionalmente, não está documentada qual é a codificação utilizada para representar a nacionalidade de falecidos que não sejam naturais do Brasil, não havendo correspondência com outras codificações de países utilizadas no SUS, como no e-SUS APS[3].

Para obter o nome do estabelecimento de saúde a partir do código CNES, é preciso ter o cuidado de buscar a tabela do CNES válida para o mesmo período da coleta de dados do SIM, caso contrário alguns registros não serão identificados. Além disso, em alguns registros foi atribuído um código CNES com todos os dígitos 9, provavelmente para associar ao valor "ignorado", embora não necessariamente a variável devesse ter um valor atribuído.

Os códigos de ocupação adotados (para falecido e, se for o caso, a mãe) não seguem fielmente a tabela oficial do Ministério do Trabalho e Emprego, usando vários códigos complementares que precisaram ser descobertos em outros formulários do SUS.

4.5 ANOMALIAS TEMPORAIS E ESTRUTURAIS

A variável Data do atestado teve, em 2024, registros com datas incompatíveis com a realidade, possivelmente por erros de transcrição ou de preenchimento.

Há potenciais problemas com a variável Tipo de ocorrência do óbito (que no sistema recebe o nome TPOBITOCOR). O Dicionário de Dados apresenta duas variáveis com esse nome e com definições diferentes. A definição alternativa e supostamente obsoleta, com informação similar à Situação gestacional, tem um conjunto de valores compatível com os dados armazenados no SIM. Isso sugere que um pós-processamento para o preenchimento das variáveis obsoletas a partir dos dados correntes está sobrescrevendo a informação original.

Além da dificuldade associada ao preenchimento com múltiplos códigos CID por linha, no SIM é acrescentado um 'X' ao final de códigos CID de três caracteres. Para fazer a integração com outras fontes de dados que usam o código CID é preciso remover esse caractere extra. Dentro de cada campo, os códigos CID são separados por um asterisco.

4.6 DETERMINANTES SOCIAIS DA SAÚDE

Observa-se que algumas das maiores fragilidades de completitude e consistência concentram-se justamente em variáveis relacionadas a determinantes sociais da saúde (Buss; Pellegrini Filho, 2007), condições reprodutivas e contextos de maior complexidade social e

[3] https://integracao.esusaps.bridge.ufsc.tech/v560/ledi/documentacao/referencias/paises.html

assistencial. Campos relacionados à escolaridade, ocupação, saúde materna e situação gestacional apresentaram inconsistências relevantes, incluindo registros de gravidez associados a indivíduos do sexo masculino e a meninas em idades muito precoces. Embora parte dessas ocorrências possa decorrer de erros de preenchimento ou limitações estruturais dos sistemas de informação, tais inconsistências também podem refletir desafios enfrentados pelos profissionais no registro de situações socialmente sensíveis, ambíguas ou complexas, incluindo contextos de vulnerabilidade social, violência e diversidade de gênero. Nesse sentido, as fragilidades observadas não dizem respeito apenas à qualidade técnica dos dados, mas também à capacidade dos sistemas e das práticas profissionais de representar adequadamente experiências humanas complexas no contexto da morte. Esses achados reforçam a necessidade de investimento contínuo em capacitação profissional, revisão terminológica e aperfeiçoamento dos instrumentos de coleta e registro em saúde.

Por fim, deve-se observar que uma futura versão eletrônica da DO poderia minimizar diversos dos problemas relatados aqui, oferecendo desde consistência básica de campos individuais até a verificação de dependências entre campos, indicando quais devem ou não ser preenchidos a partir de outros valores e reforçando consistências temporais. A DO eletrônica poderia também resolver problemas como o preenchimento inadequado das causas de morte, não apenas com a integração com a versão oficial da CID-10 (e da CID-11, quando integralmente implementada no SUS), mas também com a possibilidade de registrar, sem os limites do formulário em papel, cadeias causais mais longas, bem como mais comorbidades, o que deve ser uma consequência natural do envelhecimento da população. No entanto, o principal investimento que pode e deve ser feito para melhorar a qualidade de dados do SIM e de outros sistemas de informação de saúde no âmbito do SUS é a formação adequada dos profissionais responsáveis pela produção dos dados.

5 CONCLUSÃO

A análise exploratória dos dados do Sistema de Informação sobre Mortalidade evidenciou que a avaliação da qualidade da informação em sistemas nacionais de saúde ultrapassa a simples mensuração de completitude de campos, envolvendo também aspectos relacionados à coerência semântica, integridade estrutural, dependências lógicas e interoperabilidade entre sistemas e classificações. Os resultados revelaram padrões recorrentes de inconsistências associados tanto ao preenchimento das Declarações de Óbito quanto às próprias limitações presentes nos mecanismos de representação computacional e codificação

adotados pelo sistema. Foram identificadas situações de preenchimento incompatível com regras condicionais do formulário, ambiguidades classificatórias, registros temporalmente inconsistentes e dificuldades de integração semântica com bases externas.

Ao mesmo tempo, observou-se que o SIM constitui uma infraestrutura robusta e estratégica para vigilância epidemiológica, pesquisa científica e formulação de políticas públicas no Brasil, apresentando elevados níveis de preenchimento em diversas variáveis analisadas. Os achados sugerem que a qualidade dos dados em saúde deve ser compreendida como resultado de um ecossistema sociotécnico complexo, envolvendo documentos normativos, sistemas computacionais, classificações padronizadas e práticas humanas de registro da informação.

O objetivo de qualquer análise exploratória de dados é estabelecer uma base mais sólida para futuros estudos e, neste caso, não é diferente. Pesquisas futuras podem avaliar, por exemplo, se muitas dessas inconsistências são simples erros ou são reflexos de preconceitos em relação a segmentos mais vulneráveis da população.

Nesse contexto, o estudo reforça a importância de investimentos contínuos em capacitação profissional, revisão de instrumentos documentais, harmonização terminológica e fortalecimento das estratégias de interoperabilidade e governança de dados em saúde digital. Além disso, demonstra o potencial da Análise Exploratória de Dados como abordagem metodológica para identificação de fragilidades estruturais e compreensão de padrões informacionais em grandes bases de dados em saúde.

DECLARAÇÃO DE AUTORIA

ILMR idealizou e desenvolveu o estudo, bem como elaborou a primeira versão do manuscrito. MCBG realizou revisão crítica do conteúdo, da metodologia e da apresentação dos resultados.

AGRADECIMENTOS

Os autores agradecem o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) via processos: 406079/2023-4 Grupos Emergentes e 309698/2025-1 Produtividade em Pesquisa.

REFERÊNCIAS

BRASIL. Ministério da Saúde. DATASUS. **Transferência de arquivos**. [S.l]: Ministério da Saúde, 2026. Disponível em: https://datasus.saude.gov.br/transferencia-de-arquivos/. Acesso em: 21 maio 2026.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. **Declaração de Óbito:** Manual de instruções para preenchimento. Brasília, DF: Ministério da Saúde, 2022. Disponível em: http://bvsms.saude.gov.br/bvs/publicacoes/declaracao_obito_manual_preenchimento.pdf. Acessado: 8 maio 2026.

BUSS, Paulo Marchiori; PELLEGRINI FILHO, Alberto. A Saúde e seus Determinantes Sociais. **Physis: Rev. Saúde Coletiva**, Rio de Janeiro, vol. 17, nº 1, p. 77–93, 2007.

FUNDAÇÃO OSWALDO CRUZ. Plataforma de Ciência de Dados aplicada à Saúde (PCDaS). **Sistema de Informação sobre Mortalidade – SIM**. [S.l: s.n], 2018. Disponível em: https://pcdas.icict.fiocruz.br/conjunto-de-dados/sistema-de-informacoes-de-mortalidade-sim/. Acesso em: 21 maio 2026.

MATHERS, Colin D.; FAT, Doris Ma; INOUE, Mie; RAO, Chalapati; LOPEZ, Alan D. Counting the dead and what they died from: an assessment of the global status of cause of death data. **Bulletin of the World Health Organization**, Geneva, v. 83, n. 3, p. 171–177, 2005.

PEARSON, Ronald K. **Exploratory Data Analysis using R**. Boca Raton: CRC Press, 2018.

PORTAL BRASILEIRO DE DADOS ABERTOS. **Sistema de Informação sobre Mortalidade – SIM**. [S.l.]: Ministério da Saúde, 2022. Disponível em: https://dados.gov.br/dados/conjuntos-dados/sim-1979-2019. Acesso em: 21 maio 2026.

REBOUÇAS, Poliana; ALVES, Flavia Jôse; FERREIRA, Andrêa; MARQUES, Lays; GUIMARÃES, Nathalia Sernizon; DE SOUZA, Giesy Ribeiro; PINTO, Priscila F.P.S.; TEIXEIRA, Camila; ORTELAN, Naiá; SILVA, Natanael; ROCHA, Aline; FALCÃO, Ila; JUNIOR, Elzo Pereira Pinto; PESCARINI, Julia; PAIXÃO, Enny S.; DE ALMEIDA, Marcia Furquim; SILVA, Rita de Cassia Ribeiro; ICHIHARA, Maria Yury Travassos; BARRETO, Maurício L. Avaliação da qualidade do Sistema Brasileiro de Informações sobre Mortalidade (SIM): uma scoping review. **Ciência e Saúde Coletiva**, Rio de Janeiro, v. 30, n. 1, 2025. Disponível em: https://doi.org/10.1590/1413-81232025301.08462023. Acesso em: 21 maio 2026.

SENNA, Mônica de Castro Maia. Sistema de Informações sobre Mortalidade (SIM). *In:* **A Experiência Brasileira em Sistemas de Informação em Saúde**: Falando sobre os Sistemas de Informação em Saúde no Brasil. Brasília, DF: Editora MS, 2009. v. 2. p. 87–105.

WICKHAM, Hadley. **Tidyverse**: R packages for Data Science. [S.l]: Posit Software, 2016. Disponível em: https://tidyverse.org. Acesso em: 21 maio 2026.