A [Secretaria de Estado de Saúde](https://www.saude.mg.gov.br/coronavirus), em parceria com a [Controladoria Geral do Estado](http://cge.mg.gov.br/noticias-artigos/740-novos-dados-sobre-a-covid-19-em-mg-disponibilizados-em-formato-aberto), passou a divulgar, em abril, microdados[[1]](#endnote-1) sobre [casos notificados](http://www.transparencia.dadosabertos.mg.gov.br/dataset/casos-notificados-coronavirus), [casos confirmados](http://www.transparencia.dadosabertos.mg.gov.br/dataset/casos-confirmados-coronavirus) e [óbitos confirmados](http://www.transparencia.dadosabertos.mg.gov.br/dataset/obitos-confirmados-coronavirus) do novo coronavírus (COVID-19). Tal iniciativa foi adotada em função da decisão de se imprimir maior transparência sobre os dados dos boletins epidemiológicos e resultados das ações de governo no enfrentamento à pandemia.

No entanto, a disponibilização de acesso irrestrito a microdados traz à tona a necessidade de encontrar o ponto de equilíbrio entre a maximização da utilidade da base de dados *versus* a minimização do risco de re-identificação[[2]](#endnote-2)

Esse dilema tem sido objeto de extensa discussão no âmbito acadêmico e profissional[[3]](#footnote-1), e vem se tornando cada vez mais relevante em virtude do fortalecimento: de um lado, da política de publicação de dados abertos governamentais; e, de outro, da política de proteção de dados pessoais.

A Open Knowledge Brasil (OKBR)[[4]](#footnote-2) reconhece a necessidade do equilíbrio em sua avaliação semanal sobre a transparência dos dados da COVID-19 divulgados por governos estaduais. Sua [metodologia de avaliação](https://transparenciacovid19.ok.org.br/files/Nota_Metodologica_Transparencia_da_Covid-19V.2.pdf) observa que

*a divulgação anonimizada de dados relativos a cada caso, separadamente, é importante para a construção de um panorama mais preciso sobre o avanço do novo coronavírus no país.* ***Informações sobre sexo, idade, município, possível origem do contágio e tratamento conferido são exemplos de dados que podem ajudar a construir uma boa base de dados de monitoramento****.*

Ao mesmo tempo que defende em seu [*Toolkit* de Publicação de Microdados](https://transparenciacovid19.ok.org.br/files/Toolkit_1_microdados_basicos.pdf) que

*Dados de saúde são considerados dados sensíveis pela Lei Geral de Proteção de Dados Pessoais, e, por isso,* ***é necessário ter o máximo de cautela ao coletar e administrar esses dados****. No caso da disponibilização de dados sobre a Covid-19, a solução mais simples para mitigar os riscos à privacidade se encontra nos diferentes níveis de detalhamento e agregação que podem ser empregados aos dados coletados.* ***Quando a publicação vincula diretamente atributos de idade e sexo dos pacientes a seus respectivos locais de internação, por exemplo, isso pode facilitar a identificação de um determinado paciente, sobretudo em municípios pequenos****. Esse tipo de publicação não é incentivada pelo ITC-19. No entanto, uma opção para não expor os pacientes e nem deixar de publicar esses dados – que são todos igualmente importantes – é divulgar o quantitativo de casos por hospital, e, separadamente, publicar os Microdados de casos com as informações de idade e sexo.*

Sabe-se que a mera remoção dos atributos de identificação direta de uma pessoa, como nome e CPF, não é suficiente para afastar o risco de re-identificação. A Lei Geral de Proteção de Dados prevê inclusive que a [anonimização](http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm#art5) consiste na utilização de meios técnicos razoáveis e disponíveis no momento do tratamento, por meio dos quais **um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo.**

Em âmbito internacional, não foi possível identificar experiências análogas com relação à divulgação oficial de microdados, que divulguem, por exemplo, um subconjunto anonimizado das [variáveis coletadas](https://www.who.int/who-documents-detail/data-dictionary-for-case-based-reporting-form) pela Organização Mundial da Saúde para fins de [monitoramento da evolução dos casos de COVID-19](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/surveillance-and-case-definitions), apesar do interesse de pesquisadores sobre essas informações[[5]](#footnote-3).

Isso pode ser explicado pelo desafio técnico, legal e ético para disseminação de arquivos de microdados, como pontuado pela [International Household Survey Network (IHSN)](https://ihsn.org/dissemination-of-microdata-files). Do ponto de vista técnico, a IHSN indica alguns mecanismos para mitigação da possibilidade de reidentificação, como

*\* divulgação da amostragem, em vez do universo dos dados;*

*\* categorização de medidas de variáveis em novas categorias, especialmente quando valores extremos são raros;*

*\* combinação de duas variáveis formando uma nova variável agregada;*

*\* remoção de variáveis*

Lidamos, portanto, com um contexto em que divulgar microdados parece ser um problema geral, ao mesmo tempo que relevante e não trivial. Tomando tais aspectos em conta, bem como os [conjuntos de dados divulgados sobre a COVID-19](http://www.transparencia.dadosabertos.mg.gov.br/organization/secretaria-de-estado-de-saude), que contém dados pessoais como por exemplo:

- classificação do caso (confirmado, suspeito, descartado ou óbito) e suas datas de notificação e de atualização;

- idade;

- sexo;

- comorbidades (sim ou não);

- município de residência[[6]](#footnote-4)

Solicitamos assistência técnica especializada para, no caso concreto de divulgação de microdados da COVID-19 pelo Estado de Minas Gerais, determinar:

**\* \_Como mensurar o risco de re-identificação de tais pacientes?**

**\* Em havendo risco alto, considerado o apetite a risco da organização, quais procedimentos adicionais seriam adequados para minimizá-lo?**

**\* Em havendo risco baixo, considerado o apetite a risco da organização, como avaliar o risco adicional incorrido com a divulgação de variáveis adicionais?**

1. Microdados são informações sobre cada unidade de observação de um fenômeno ou evento. No caso em questão, cada linha do arquivo eletrônico contém uma observação de um paciente acometido com o coronavírus, sendo que as características coletadas estão representadas nas colunas subsequentes às do ID do paciente (valores de cada variável: idade, gênero, etc) [↑](#endnote-ref-1)
2. Possibilidade de utilizar outras variáveis dos microdados para identificá-los, de fato. No caso em questão, seria o risco de associação de variáveis como idade, gênero, município de residência, evento de internação e/ou comorbidade a outras bases de dados disponíveis, para descobrir o nome ou identificação do paciente. [↑](#endnote-ref-2)
3. [^garfinkel2016]: De-Identifying Government Datasets (2nd Draft) [↑](#footnote-ref-1)
4. Organização sem fins lucrativos de promoção do conhecimento livre com enfoque em dados abertos [↑](#footnote-ref-2)
5. [Epidemiological data from the COVID-19 outbreak, real-time case information](https://www.nature.com/articles/s41597-020-0448-0) [↑](#footnote-ref-3)
6. Existem 853 municípios no estado de Minas Gerais, [dos quais 666 possuem menos de 20 mil habitantes](https://pt.wikipedia.org/wiki/Lista_de_munic%C3%ADpios_de_Minas_Gerais_por_popula%C3%A7%C3%A3o) [↑](#footnote-ref-4)