cadernos Saúde Coletiva
ISSN 2358-291X (Online)

ARTIGO ORIGINAL

Ciência de dados e big data: o que isso significa para estudos populacionais e da saúde?

Data science and big data: what do these terms mean for population and health related studies?

Raphael de Freitas Saldanha[1], Christovam Barcellos[1], Marcel de Moraes Pedroso[1]

[1]*Fundação Oswaldo Cruz (Fiocruz) - Rio de Janeiro (RJ), Brasil.*

Como citar: Saldanha RF, Barcellos C, Pedroso MM. Ciência de dados e big data: o que isso significa para estudos populacionais e da saúde?. Cad Saúde Colet, 2021;29(esp.):51-58. https://doi.org/10.1590/1414-462X202199010305

Resumo

Introdução: O termo "*big data*" no ambiente acadêmico tem deixado de ser uma novidade, tornando-se mais comum em publicações científicas e em editais de fomento à pesquisa, levando a uma revisão profunda da ciência que se faz e se ensina. **Objetivo**: Refletir sobre as possíveis mudanças que as ciências de dados podem provocar nas áreas de estudos populacionais e de saúde. **Método**: Para fomentar esta reflexão, artigos científicos selecionados da área de *big data* em saúde e demografia foram contrastados com livros e outras produções científicas. **Resultados**: Argumenta-se que o volume dos dados não é a característica mais promissora de *big data* para estudos populacionais e de saúde, mas a complexidade dos dados e a possibilidade de integração com estudos convencionais por meio de equipes interdisciplinares são promissoras. **Conclusão**: No âmbito do setor de saúde e de estudos populacionais, as possibilidades da integração dos novos métodos de ciência de dados aos métodos tradicionais de pesquisa são amplas, incluindo um novo ferramental para a análise, monitoramento, predição de eventos (casos) e situações de saúde-doença na população e para o estudo dos determinantes socioambientais e demográficos.
Palavras-chave: saúde pública; demografia; ciência de dados; *big data*.

Abstract

Background: The term big data is no longer new in the academic environment and has become more common in scientific publications and research grants, leading to a profound revision of the way science is being made and taught. **Objective**: To reflect on the possible changes that data science can induce in population and health related studies. **Method**: To foster this debate, scientific articles selected from the big data field in health and demography were contrasted with books and other scientific productions. **Results**: It is argued that volume is not the most promising characteristic of big data for population and health related studies, but rather the complexity of data and the possibilities of integration with traditional studies by means of interdisciplinary teams. **Conclusion**: In population and health related studies, the possibilities of integration between new and traditional methods are broad, and include new toolboxes for analysis, monitoring, prediction of events (cases) and health-disease processes in the population, and for the study of sociodemographic and environmental determinants.

Keywords: public health; demography; data science; big data.

Trabalho realizado no Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT), Fundação Oswaldo Cruz (Fiocruz) – Rio de Janeiro (RJ), Brasil.
Correspondência: Raphael de Freitas Saldanha. E-mail: raphael.saldanha@icict.fiocruz.br
Fonte de financiamento: Fundação Oswaldo Cruz (Fiocruz) – Rio de Janeiro (RJ), Brasil.
Conflito de interesses: nada a declarar.
Recebido em: Ago. 15, 2019. Aprovado em: Jan. 08, 2020

cadernos Saúde Coletiva

INTRODUÇÃO

O termo "*big data*" no ambiente acadêmico tem deixado de ser uma novidade, tornando-se mais comum em publicações científicas[1] e em editais de fomento à pesquisa[2 1]. Departamentos de universidades e centros de pesquisa internacionais e nacionais têm redesenhado suas ementas ou criado programas e disciplinas para atender à demanda de formação em ciência de dados e *big data*[3]. Passado o *hype* da novidade, resta refletir sobre as mudanças necessárias e seus impactos nas áreas de estudos populacionais e de saúde.

Os sistemas de informação de saúde e os recenseamentos, inquéritos e outras pesquisas demográficas são as principais fontes de dados para o conhecimento das dinâmicas populacionais. Tais pesquisas já produzem um respeitoso volume de dados, introduzindo complexidades para a sua análise. Então, já éramos *big data*? Ou *big data* se define não só pela quantidade de dados em análise?

Diversas propostas para definição do que vem a ser *big data* (ou sua vertente científica e acadêmica denominada "ciência de dados") se apresentam, acompanhando, naturalmente, a evolução de um campo novo, com a absorção de suas rápidas mudanças e novas aplicações. Contudo, parece haver uma convergência: *big data* não diz respeito, unicamente, a grandes bases de dados. As implicações da correção dessa característica para estudos populacionais e de saúde são importantes. O volume e a velocidade da produção de dados dessas áreas em comparação aos campos das ciências naturais, como física e biologia, são menores, mesmo comparando-se todos os microdados de todas as pesquisas demográficas e dos sistemas de informação em saúde.

Nesse contexto, o termo "ciência de dados" vem se consolidando como um campo de convergência tecnológica, científica, acadêmica, filosófica e pragmaticamente interdisciplinar, formado, basicamente, por cientistas da computação, matemáticos, estatísticos e pesquisadores com conhecimento substantivo do problema em análise – como os médicos e sanitaristas no caso na saúde –, mas é possível incluir aqui grupos de pesquisas em ciência de dados que contam com biólogos, geneticistas, economistas, financistas, geógrafos, advogados, historiadores, entre outros profissionais em suas equipes.

A definição utilizada em nosso grupo de pesquisa[4 2] para esse domínio em construção é: "Ciência de Dados é um campo de estudo que se destaca pela capacidade de auxiliar a descoberta de informação útil a partir de grandes ou complexas bases de dados, bem como a tomada de decisão orientada por dados"[3].

Desta forma, entende-se que *big data* é um dos aspectos do campo da ciência de dados que trata de outros aspectos, como estratégias para extração, transformação e carga dos dados, modelagem, construção e avaliação de algoritmos descritivos e preditivos, visualização de grandes quantidades de dados e *deploy* dos modelos em ambientes de produção para a tomada de decisão, entre outros. O que importa na definição de *big data* não é o volume ou mesmo a

[1] Levantamento não sistemático realizado pelos autores em junho de 2019 no *Web of Science*, via Portal de Periódicos CAPES, utilizando os descritores "big data" e "health" nos campos "título" e "assunto" e com o filtro de tópico "big data" ativado, obtendo o retorno de 3.728 artigos revisados por pares em revistas científicas indexadas, sobretudo a partir de 2012.

[2] Iniciativa recente foi a chamada conjunta do *Grand Challenges Explorations*, voltada exclusivamente, e pela primeira vez, a pesquisadores brasileiros, resultado da parceria entre o Ministério da Saúde (MS), o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), o Conselho Nacional das Fundações Estaduais de Amparo à Pesquisa (CONFAP), as Fundações Estaduais de Amparo à Pesquisa (FAPs) e a Fundação Bill & Melinda Gates (FBMG). Ver detalhes em Global Grand Challenges[1].

[3] Alguns exemplos internacionais e nacionais: Harvard Data Science Initiative; LSE: Data Science; University of California: MIT: Data Science and Big Data Analytics Making Data-Driven Decisions; Master of Information and Data Science; University of Denver: Master of Science in Data Science; Syracuse University: Master in Data Science; Columbia University: Master of Science in Data Science; Universidade de Lisboa: Ciência de Dados; USP: São Paulo School of Advanced Science on Learning from Data; UNICAMP: Difusão em Ciência de Dados; PUC-RIO: Ciência de Dados; PUC-Minas: Ciência de Dados e Big Data; CEFET/RJ: Ciência da Computação com ênfase em Ciência de Dados; Icict/Fiocruz: Ciência de Dados aplicada à Saúde; Farmanguinhos/Fiocruz: Big Data em Saúde; LNCC: Programa Multidisciplinar de Pós-Graduação do Laboratório Nacional de Computação Científica.

[4] Grupo de pesquisa em Ciência de Dados aplicada à Saúde, cadastrado no CNPq e certificado pela Fiocruz[2].

velocidade da produção de dados, mas a complexidade estrutural desses dados (variedade) e o poder computacional necessário para analisá-los integralmente.

Essa complexidade deriva não só da multiplicidade de fontes de dados e de variáveis inter-relacionadas, mas também da estrutura sociodemográfica contemporânea, imbricada por fatores econômicos, culturais e políticos. A necessidade de se atribuir grandes categorias sociais às populações tem, de fato, dominado o debate tanto das ciências sociais quanto do discurso político, que costuma utilizar termos simplificadores para designar a classe pobre (descamisados, excluídos, vulneráveis, despossuídos, miseráveis, *sans-culottes*, proletários, entre outros), a classe média (pequena burguesia, emergentes, intelectuais, servidores públicos etc.) e a mais rica também (elite, patronato, classe alta etc.).

Na sociologia, também os recortes teóricos são produzidos segundo concepções ideológicas, com marcadas contribuições de escritos clássicos de Marx, Durkheim e Weber[4]. Se, para Marx, a estrutura social capitalista é resultado das relações de produção, que gera uma sociedade de classes antagônicas e complementares, para Weber a estratificação social deriva da distribuição desigual de riqueza, prestígio e poder. Mais recentemente, Bourdieu introduziu o conceito de capital cultural, identificando distinções de classe pelos seus valores, hábitos e padrões de consumo de bens e serviços[5], deslocando o debate sobre inclusão/exclusão social da esfera da produção para a esfera da circulação e do consumo.

Enquanto alguns desses conceitos podem ser operacionalizados para a coleta e a análise de dados sociodemográficos e traduzidos em perguntas objetivas que irão compor os questionários de censos, de inquéritos e os registros de saúde, como a posse de bens, cor e raça, sexos (mas não gênero), nível educacional e de ocupação, outros são de difícil apreensão, como a detenção de meios de produção, a identidade de classe e etnicidade, o prestígio social e a inserção nos circuitos econômicos, culturais e de poder.

Qualquer que seja a abordagem teórica sobre a estrutura social, é evidente que essa estrutura e estratificação constituem um fenômeno emergente da sociedade de classes, e a posição social de cada indivíduo é definida pela sua inserção política, econômica e simbólica nessa sociedade. No caso dos estudos quantitativos, esses componentes somente podem ser identificados *a posteriori*, isto é, pela análise integrada e multidimensional de uma variedade e quantidade de dados que captam os diversos aspectos dessa inserção do indivíduo na sociedade[6]. Esse tipo de análise pode ser viabilizado pelas técnicas estatísticas, pelo manuseio criativo das fontes de dados tradicionais, pela incorporação de novas fontes de dados e pela infraestrutura computacional atualmente disponível.

Não por acaso, os primeiros censos demográficos geraram tabelas em geral bivariadas, contendo a quantidade de habitantes segundo o sexo, estado civil, nacionalidade, religião, grau de instrução e profissão, com uma marcada diferença entre pessoas "escravas e livres", conforme ilustrado na figura 1. Novos censos, realizados principalmente após 1970, introduziram diversas outras variáveis e instrumentos de coleta de dados[7]. A partir desses novos dados, foi possível criar outras formas de agregação de dados e realizar análises estatísticas multivariadas.

Figura 1. Dados da "Província do Rio Grande do Norte", censo de 1872. Fonte: IBGE[7]

cadernos Saúde Coletiva

Mas a complexidade da sociedade brasileira, entre outros fatores, exige novas estratégias de categorização e análise. As análises estatísticas convencionais tendem a identificar correlações ou associações entre variáveis e grandes padrões de semelhança de grupos populacionais. As técnicas de *data mining* e a diversidade de variáveis atualmente disponíveis permitem detectar disjunções, isto é, combinações entre variáveis que conformam grupos minoritários, militantes, projetos de vida alternativos e desconexões, que fogem aos padrões sociais hegemônicos e podem ser capturados mediante a "mineração" de fontes alternativas de informação, como as redes sociais e outras formas de interação entre usuários via aplicativos.

Estão assentadas todas as condições tecnológicas para se trabalhar com o par de conceitos dialeticamente ligados de desigualdade e diferença, ou inclusão e exclusão[8], pois são essas categorias analíticas que podem explicar a vulnerabilização e a marginalização de grupos sociais, bem como elucidar os diferentes graus de exposição aos determinantes socioambientais do processo saúde-doença intragrupos populacionais.

A BUSCA DE PADRÕES E DIVERGÊNCIAS

Nas ciências da natureza, mais dados significam, basicamente, maiores computadores ou necessidade de métodos que consigam lidar de modo eficiente com mais dados. Até certo momento, as ciências sociais também seguiam essa lógica.

Tradicionalmente, após a definição da pergunta de estudo e da área de estudo, faz-se um recorte das bases de dados, levando-se apenas as variáveis de interesse já previamente sugeridas por levantamentos bibliográficos ("pela literatura"). Essas variáveis são confrontadas em um modelo estatístico de verificação de hipóteses *a priori* que busca responder à pergunta inicial do estudo (*hypothesis-driven approaches*).

A inovação metodológica promissora na utilização de técnicas de *big data* nas ciências sociais e, particularmente, na saúde pública é permitir se fazer uma ciência *a posteriori*; é procurar padrões, localizar associações, visualizar a complexidade dos fenômenos, prever desfechos em saúde sem considerar previamente hipóteses formuladas *a priori*, predizer comportamentos com precisão sem necessariamente partir de uma sustentação teórica ou clínica; é transitar de "theory-driven approach" ou "hypothesis-driven approach" para "data-driven approach". Conforme afirma Bohon[9:5]:

> Nossos problemas também requerem novos modos de pensar sobre os dados que temos e sobre novos métodos para analisar e visualizar estes dado.s [...] Não podemos simplesmente escalonar, precisamos mudar quase tudo.

Trabalhar com *big data* em estudos populacionais e de saúde pública não significa, necessariamente, trabalhar com muitos dados, mas diz respeito a alterar, de maneira profunda, o modo de se fazer pesquisa, o que leva à necessidade de reformulação dos currículos de graduação e pós-graduação. Faz-se necessário ir além da mudança de nomes de disciplinas e departamentos, para uma revisão profunda da ciência que se faz e se ensina.

Aos críticos e céticos quanto à viabilidade da adoção disruptiva da abordagem *data-driven* em estudos populacionais e de saúde, cabe salientar que a discussão epistemológica não pode ser baseada em visões dicotômicas da ciência do tipo "qualitativo ou quantitativo" encaradas, frequentemente, como abordagens antagônicas. A proposta aqui defendida é de construir caminhos para uma ciência aberta[5] [10], criativa, inovadora, que possa adotar, de forma simultânea, métodos mistos (qualitativos e quantitativos) e que possa ser guiada por procedimentos híbridos (*hypothesis* e *data-driven*)[11-14].

Há uma evidente associação entre as variáveis socioeconômicas coletadas pelos sistemas de informação de saúde. A disponibilização de dados e o interesse retomado da epidemiologia pela busca da equidade em saúde fizeram crescer o número de estudos sobre desigualdades

[5] Destaque para a adesão crítica e estratégica da Fiocruz ao movimento global de "Ciência Aberta", que, em linhas gerais, propõe tornar a pesquisa científica acessível a todos. Na prática, significa eliminar obstáculos artificiais, especialmente os editoriais, legais e econômicos, à livre circulação do conhecimento científico. É possível acessar o conjunto de iniciativas em Fundação Oswaldo Cruz[10].

sociais e de saúde[15]. Renda, educação, etnicidade e ocupação se manifestam na sociedade brasileira como uma conjunção de fatores que podem acarretar melhores ou piores condições de vida e de saúde.

A adoção de grandes categorias e o uso de indicadores compostos e sintéticos têm sido empregados para a detecção dessas desigualdades[15]. No entanto, essas abordagens não são capazes de responder sobre os efeitos na saúde de contextos particulares de risco, como grupos indígenas com crenças, hábitos e atitudes tradicionais que produzem perfis epidemiológicos característicos e os afastam dos padrões urbanos e das categorias classicamente usadas para estudos sobre desigualdades. O que dizer, por exemplo, sobre os idosos urbanos, que podem possuir uma renda minimamente necessária para seu sustento, mas se encontram isolados por causa de suas condições clínicas e familiares? Teriam eles um perfil diferenciado de risco? Mais que isso, como vivem as mulheres pobres e negras da periferia das grandes cidades, submetidas a uma sobreposição de condições adversas de violência, dificuldades de acesso a bens e serviços e sujeitas a discriminações na cidade e nas instituições? Apontar os perfis particulares de risco de cada grupo populacional pode ser um importante passo para a construção de políticas de saúde inclusivas e mais adequadas a cada grupo.

Além disso, é importante ressaltar que pessoas com perfis sociodemográficos semelhantes, mas que moram e circulam em lugares diferentes, podem apresentar perfis epidemiológicos discrepantes. O território em que as pessoas moram define grande parte das condições de vida, da produção de doenças, de acesso aos serviços de saúde e da organização social local, o que permite estabelecer laços de solidariedade e compartilhar recursos para sua proteção[16]. A restituição dos dados coletados em censos ou registros de saúde ao seu lugar de origem, realizada atualmente por procedimentos de georreferenciamento, permite resgatar o contexto em que se produzem os riscos e a intensidade e frequência das exposições da população ao determinantes socioambientais, mas igualmente onde se devam promover o estabelecimento de sistemas de proteção social. Nesse sentido, as análises de dados por técnicas de *big data* não podem prescindir de informações geográficas que permitem complementar e contextualizar eventos de saúde-doença que se expressam sempre no nível individual, com dados sobre o ambiente, as condições socioeconômicas, a presença de instituições e redes de apoio, que são características subjacentes ao território.

BIG DATA EM SAÚDE, ACESSO A DADOS E DIREITO AO SIGILO

A forma como se lida com os dados também deve ser alterada. Tradicionalmente, pesquisas sobre mortalidade usam, quase que de forma exclusiva, dados do Sistema de Mortalidade; pesquisas sobre migração usam dados do Censo; pesquisas sobre orçamento e renda buscam dados de inquéritos específicos. Contudo, a complexidade dos fenômenos humanos, da sua saúde e comportamento transita livremente por essas bases de dados e as transborda. Para procurar dar conta dessa complexidade, o mínimo a fazer é procurar trabalhar com esses dados de modo integrado e interoperável.

Dados de mortalidade, natalidade, notificação de doenças, internações hospitalares e atendimentos ambulatoriais, dentre outros, são captados por sistemas de informação específicos que registram eventos em saúde, desde o nascimento até o óbito em diversos estabelecimentos de saúde ou fora destes, em diferentes ocasiões. Conectar esses eventos (e dados) permitiria traçar o percurso de pessoas dentro do sistema de saúde, além de definir a história clínica de cada pessoa e suas situações de saúde e de doença. Uma mesma pessoa pode ser portadora de HIV, sofrer violência e ser internada em uma unidade de emergência. É importante que os serviços de saúde sejam informados sobre essa combinação de fatores para poder prestar atenção médica e psicossocial adequada e oportuna. Além disso, o conjunto de histórias individuais permite estimar a prevalência de determinadas doenças crônicas, de sobrevida e identificar suas comorbidades.

Iniciativas importantes estão, por meio de métodos criativos, mistos, híbridos e intensivos em computação científica, buscando estimar dados e indicadores de saúde-doença por meio de dados públicos individualizados, porém não identificados.

cadernos Saúde Coletiva

A revista Lancet, em 2016, publicou um editorial intitulado "*GBD 2015: from big data to meaningful change*", em que apresentava aos leitores alguns dos principais achados do estudo, as estratégias metodológicas para a construção das estimativas e a publicação dos resultados do número especial "*Global Burden of Disease Study 2015*", uma parceria entre o *Institute for Health Metrics and Evaluation* (IHME) e a Lancet[17].

Em nossa opinião, a iniciativa *Global Burden of Disease* (GBD) é, atualmente, o esforço (humano e computacional) mais abrangente em estudos populacionais, epidemiológicos e de saúde que utiliza métodos híbridos, mistos e *big data* aplicados à morbimortalidade das principais doenças, agravos e fatores de risco para a saúde-doença em níveis global, nacional e regional. Disponibiliza os resultados dos estudos, publiciza as metodologias, fontes de dados utilizados para a construção de suas principais métricas – *years of life lost* (YLLs), *years lived with disability* (YLDs), e a combinação dos dois últimos *disability-adjusted life-years* (DALYs) – e suas tendências desde 1990, bem como ferramentas para visualizações interativas desses resultados, permitindo, desta forma, que pesquisadores sejam estimulados a fazer comparações entre as populações e buscar compreender os desafios para a saúde pública decorrentes dessas tendências. No estudo de 2017, foram avaliadas 359 causas de morbimortalidade prematuras, incidência e prevalência de doenças e anos vividos com incapacidade para 195 países e territórios. Para ter acesso aos estudos e publicações derivadas dos GBDs 2017, 2016, 2015, 2013 e 2010, deve-se consultar The Lancet[18].

No Brasil, um acordo em 2015 entre o Ministério da Saúde (MS), a Universidade Federal de Minas Gerais (UFMG) e o IHME, da Universidade de Washington, resultou no Projeto GBD Brasil, que visava constituir uma rede de colaboradores, com participação de pesquisadores brasileiros e técnicos do MS, com o objetivo de dar apoio metodológico e avaliar as estimativas do estudo GBD em nível subnacional, bem como compilar e analisar a carga de doenças no país e nos estados brasileiros. Os primeiros resultados do estudo no Brasil, no âmbito desse projeto, foram publicados em 2017 no suplemento da Revista Brasileira de Epidemiologia. É possível acessar esse suplemento em Revista Brasileira de Epidemiologia[19].

Em 2018, o Projeto GBD Brasil publicou o estudo "*Burden of disease in Brazil, 1990-2016: a systematic subnational analysis for the Global Burden of Disease Study 2016",* que consolida, estima e analisa mudanças políticas, econômicas e epidemiológicas que afetaram o processo saúde-doença no país. Os resultados do estudo apresentaram estimativas do GBD 2016 para expectativa de vida ao nascer, esperança de vida saudável, morbimortalidade específicas por causas (333), perda de saúde em razão de morte ou incapacidade e fatores de risco para o Brasil, seus 26 estados e o Distrito Federal, de 1990 a 2016[20].

Criado em 2017, o Laboratório de *Big Data* e Análise Preditiva em Saúde (LABDAPS) da Faculdade de Saúde Pública da Universidade de São Paulo (FSP/USP) tem o objetivo de desenvolver pesquisas que auxiliem na melhoria da atenção à saúde no Brasil. Os pesquisadores do laboratório trabalham na aplicação e no desenvolvimento de métodos de inteligência artificial (*machine learning*) a problemas importantes da área da saúde, como a análise de impacto de políticas públicas de saúde, a melhoria da qualidade da informação de saúde e a predição da ocorrência de doenças e óbitos (mais detalhes e publicações em[21]).

A Plataforma de Ciência de Dados aplicada à Saúde (PCDaS) do Laboratório de Informação em Saúde (LIS) do Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT) da Fundação Oswaldo Cruz (Fiocruz), criada em 2016 em parceria com o Laboratório Nacional de Computação Científica (LNCC) do Ministério da Ciência, Tecnologia, Inovações e Comunicações (MCTIC), é um projeto de pesquisa e desenvolvimento tecnológico que tem como objetivo principal disponibilizar serviços tecnológicos e computação científica para armazenamento, gestão, análise, visualização e disseminação de grandes quantidades de dados de saúde e seus determinantes socioambientais para pesquisadores, docentes e discentes de instituições de ensino e pesquisa, bem como gestores governamentais (mais informações em[3]). Por outro lado, a vinculação (*linkage*) de dados socioeconômicos com os diversos eventos de saúde individuais pode abrir a empresas privadas (seguros de saúde ou empregadores) ou golpistas a oportunidade de identificar pessoas com maior risco de adoecer e morrer, ameaçar, estigmatizar, excluir ou até chantagear essas pessoas. Mesmo o comportamento de uma

pessoa nas mídias sociais pode ser capturado e tratado de modo a delinear perfis de risco e probabilidades de adoecimento por meio de algoritmos[6 22], que constituem informações de grande valor de mercado[23]. Em países com sólida tradição democrática, como a Suécia, dispõe-se de uma grande massa de dados sociodemográficos e de saúde que são *lincados* por meio de um código de identificação pessoal, mantidos anonimizados em instituições de governo e disponibilizados para pesquisadores e planejadores de políticas públicas[24]. A preocupação com o uso indevido de dados individuais não é, portanto, necessariamente uma particularidade de governos autoritários, mas de sociedades modernas, nas quais técnicos capacitados podem acessar dados que se encontram nas nuvens ou em servidores públicos. Nesse contexto, a necessidade de *linkage* entre bases de dados torna-se imprescindível para a pesquisa e os tomadores de decisão, e com ela, as preocupações sobre confidencialidade e segurança[7 25].

O conservadorismo das instituições e a noção de propriedade feudal sobre "seus" dados e resultados são questões que costumam se esconder sob os princípios nobres e necessários da ética em pesquisa. Assimilar tecnologias de *big data* não significa abandonar esses princípios, mas confrontar a necessidade de completo acesso aos dados e uso deles com as restrições que são de ordem técnica, e não de ordem patrimonialista ou política.

DESAFIOS E PERSPECTIVAS FUTURAS

O uso de técnicas de *big data* em saúde pública apresenta condições para a superação de modelos simplistas de classificação de riscos e identificação de desigualdades. Novas categorias analíticas podem ser buscadas por meio de algoritmos que ajudem a definir, com maior precisão, grupos e situações de risco e vulnerabilidade. A rápida transformação sociodemográfica do Brasil exige que se analise seu impacto sobre a saúde da população sob a perspectiva de grupos socioespaciais particulares[26], o que inclui o envelhecimento e a urbanização da população, as novas formas de exclusão, que não são, necessariamente, determinadas pela renda, e a permanência de situações históricas de segregação e produção de condições adversas de vida, saúde e doença.

No âmbito do setor de saúde e de estudos populacionais, não é difícil imaginar as possibilidades da abordagem com técnicas de ciência de dados e *big data* para análise, monitoramento, predição de eventos (casos) e situações de saúde-doença na população, bem como a associação destes com seus determinantes socioambientais e demográficos.

Por outro lado, observa-se no Brasil um cenário revolto no tocante à formulação de políticas públicas e dados abertos. São agora comuns os ataques a pesquisas essenciais, como o Censo Demográfico e DETER, tentativas de retrocesso à divulgação de dados, acesso à informação pública e captura de instituições públicas detentoras de dados estratégicos por interesses privados. Enquanto a ciência avança, com a incorporação de novos métodos e possibilidades, faz-se necessário resguardar no campo cívico e político a preservação da autonomia e capacidade de inovação dos meios e modos de se fazer ciência.

REFERÊNCIAS

1. Global Grand Challenges. Grand Challenges Explorations - Brazil: Data Science Approaches to Improve Maternal and Child Health in Brazil [Internet]. 2019 [citado em 2019 ago 15]. Disponível em: https://gcgh.grandchallenges.org/challenge/grand-challenges-explorations-brazil-data-science-approaches-improve-maternal-and-child

2. Diretório dos Grupos de Pesquisa no Brasil – Lattes. Ciência de Dados aplicada à Saúde [Internet]. 2019 [citado em 2019 ago 15]. Disponível em: http://dgp.cnpq.br/dgp/espelhogrupo/4230691756969719

[6] Para ver a descrição interessante de aplicação de modelo de *machine learning* para detecção precoce de depressão em tweets, deve-se consultar[22].

[7] Destaque para a iniciativa do Centro de Integração de Dados e Conhecimentos para Saúde (Cidacs/Fiocruz/Bahia), que conduz estudos e pesquisas com base em projetos interdisciplinares originados na vinculação de grandes volumes de dados para ampliar o entendimento dos determinantes e das políticas sociais e ambientais sobre a saúde da população. Ver detalhes em[25].

3. Fundação Oswaldo Cruz. Plataforma de Ciência de Dados aplicada à Saúde - PCDaS [Internet]. 2019 [citado em 2019 jun 15]. Disponível em: https://bigdata.icict.fiocruz.br/
4. Giddens A. Capitalism and modern social theory: an analysis of the writings of Marx, Durkheim and Max Weber, Cambridge: Cambridge University Press; 1985.
5. Bourdieu P. Poder simbólico. São Paulo: Bertrand; 2002.
6. Lebaron F. How Bourdieu "quantified" Bourdieu: the geometric modelling of data. Quantifying Theory: Pierre Bourdieu. USA: Springer; 2009.
7. Instituto Brasileiro de Geografia e Estatistica. Memórias IBGE [Internet]. 2019 [citado em 2019 ago 15]. Disponível em: https://memoria.ibge.gov.br/sinteses-historicas/historicos-dos-censos/censos-demograficos.html
8. Canclini NG. Diferentes, desiguais e desconectados. Rio de Janeiro: UFRJ; 2009.
9. Bohon SA. Demography in the Big Data Revolution: changing the culture to forge new frontiers. Popul Res Policy Rev. 2018;37(3):323-41. http://dx.doi.org/10.1007/s11113-018-9464-6.
10. Fundação Oswaldo Cruz. Ciência Aberta [Internet]. 2019 [citado em 2019 ago 15]. Disponível em: https://portal.fiocruz.br/ciencia-aberta
11. Shmueli G. To Explain or To Predict? Statistical Science. Statistical Science; 2010. http://dx.doi.org/10.2139/ssrn.1351252.
12. Creswell J. Projeto de pesquisa: métodos qualitativo, quantitativo e misto. 4. ed. São Paulo: Bookman; 2010.
13. Chiavegatto ADP Fo. Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. Epidemiol Serv Saude. 2015;24(2):325-32. http://dx.doi.org/10.5123/S1679-49742015000200015.
14. Elliott KC, Cheruvelil KS, Montgomery GM, Soranno PA. Conceptions of good science in our data-rich world. Bioscience. 2016;66(10):880-9. http://dx.doi.org/10.1093/biosci/biw115. PMid:29599533.
15. da Silva JB, Barros MB. Epidemiologia e desigualdade: notas sobre a teoria e a história. Rev Panam Salud Publica. 2002;12(6):375-83. http://dx.doi.org/10.1590/S1020-49892002001200003. PMid:12690724.
16. Rojas LI. Geografía y salud: temas y perspectivas en América Latina. Cad Saude Publica. 1998;14(4):701-11. https://doi.org/10.1590/S0102-311X1998000400012.
17. The Lancet. GBD 2015: from big data to meaningful change. Lancet. 2016;388(10053):1447. http://dx.doi.org/110.1016/S0140-6736(16)31790-1.
18. The Lancet. Global Burden of Disease [Internet]. 2019 [citado em 2019 ago 15]. Disponível em: https://www.thelancet.com/gbd
19. Associação Brasileira de Pós-Graduação em Saúde. Revista Brasileira de Epidemiologia [Internet]. São Paulo: Associação Brasileira de Pós-Graduação em Saúde. 2017;20(Supl 1). [citado em 2019 ago 15]. Disponível em: https://www.scielosp.org/toc/rbepid/2017.v20suppl1/
20. GBD 2016 Brazil Collaborators. Burden of disease in Brazil, 1990–2016: a systematic subnational analysis for the Global Burden of Disease Study 2016. Lancet. 2018;392(10149):760-75. http://dx.doi.org/10.1016/S0140-6736(18)31221-2. PMid:30037735.
21. Laboratório de Big Data e Análise Preditiva em Saúde [Internet]. 2019 [citado em 2019 ago 15]. Disponível em: https://sites.google.com/view/labdaps
22. Martinez VR. A machine learning approach for the detection of depression and mental illness in Twitter [Internet]. Medium; 2019 [citado em 2019 ago 15]. Disponível em: https://medium.com/datadriveninvestor/a-machine-learning-approach-for-detection-of-depression-and-mental-illness-in-twitter-3f3a32a4df60
23. Vayena E, Salathé M, Madoff LC, Brownstein JS. Ethical Challenges of Big Data in Public Health. PLOS Comput Biol. 2015;11(2):e1003904. http://dx.doi.org/10.1371/journal.pcbi.1003904. PMid:25664461.
24. Cnudde P, Rolfson O, Nemes S, Kärrholm J, Rehnberg C, Rogmark C, et al. Linking Swedish health data registers to establish a research database and a shared decision-making tool in hip replacement. BMC Musculoskelet Disord. 2016;17(1):414. http://dx.doi.org/10.1186/s12891-016-1262-x. PMid:27716136.
25. Centro de Integração de Dados e Conhecimentos para Saúde [Internet]. 2019 [citado em 2019 ago 15]. Disponível em: https://cidacs.bahia.fiocruz.br/
26. Barcellos C, Sabroza PC, Peiter P, Rojas LI. Organização espacial, saúde e qualidade de vida: análise espacial e uso de indicadores na avaliação de situações de saúde. Inf Epidemiol SUS. 2002;11(3):129-38. http://dx.doi.org/10.5123/S0104-16732002000300003.