Uso e Proteção de Dados Pessoais na Pesquisa Científica

Use and Protection of Personal Data in Scientific Research

MAURÍCIO LIMA BARRETO[1]

Pesquisador (Especialista) do Centro Integrado de Dados e Conhecimento para a Saúde – Fiocruz/Bahia, Médico (UFBa), Mestre em Saúde Comunitária (UFBa), Ph.D. em Epidemiologia (LSHTM-U de Londres), Professor titular aposentado em Epidemiologia do ISC/UFBA, Professor Permanente do PPGSC-UFBa.

BETHANIA DE ARAUJO ALMEIDA[2]

Graduação e Mestrado em Ciências Sociais e Doutorado em Saúde Coletiva (UFBA). Atua no Centro de Integração de Dados e Conhecimentos para a Saúde (Cidacs) da Fundação Oswaldo Cruz e no Grupo de Trabalho em Ciência Aberta da mesma instituição.

DANILO DONEDA[3]

Bacharel em Direito pela Universidade Federal do Paraná (1995), Mestre (1999) e Doutor em Direito pela Universidade do Estado do Rio de Janeiro (2004), Professor visitante na Faculdade de Direito da Universidade do Estado do Rio de Janeiro. Foi Coordenador-Geral de Estudos e Monitoramento de Mercado na Secretaria Nacional do Consumidor do Ministério da Justiça. Foi pesquisador visitante na Università degli Studi di Camerino e na Autorità Garante per la Protezione dei Dati Personali, ambas na Itália.

RESUMO: A pesquisa científica e, em particular, a pesquisa em saúde utilizam dados pessoais e recorrem a marcos éticos para disciplinar a utilização desses dados e a redução dos riscos potenciais sem comprometer a qualidade do trabalho e a relevância do resultado. A utilização de bancos de dados das mais diversas naturezas vem se mostrando uma alternativa para o desenvolvimento da pesquisa, o que intensifica a necessidade que se trate conjuntamente o debate sobre aspectos éticos do uso de dados pessoais em pesquisa com os seus aspectos legais, considerando que os mais recentes marcos regulatórios de proteção de dados trazem regras específicas referentes à pesquisa científica.

PALAVRAS-CHAVE: Uso de dados pessoais; ética e proteção de dados; pesquisa científica.

ABSTRACT: Scientific research and, particularly, health research, demands the treatment of personal data and uses ethical frameworks to regulate the use of these data as well as to reduce its potential

1 Orcid: <https://orcid.org/0000-0002-0215-4930>.
2 Orcid: <https://orcid.org/0000-0001-8918-2661>.
3 Orcid: <https://orcid.org/0000-0001-9535-3586>.

risk, while trying not to compromise the quality of the research and the relevance of its results. The use of diverse databases are an important and unavoidable alternative to the development of scientific research, fostering the need to unify the current debates on the ethical aspects of the use of personal data in scientific research with the one regarding its legal aspects, which is even more important with the advent of new data protection legislation that tackles the issue of scientific research.

KEYWORDS: Use of personal data; ethics and data protection; scientific research.

INTRODUÇÃO

A ciência é um sistema social particular orientado por um conjunto institucionalizado de crenças, princípios e normas que conformam papéis sociais, formas de produção de conhecimento, sistemas de avaliação e recompensas, entre outros aspectos que possuem bases históricas, sociais e culturais (Merton, 2013; Khun, 2009).

Entre as normas e os valores que orientam a prática científica estão compromissos com a ética e a integridade na pesquisa, respectivamente relacionados a valores universais e a compromissos de conduta que devem ser respeitados perante os participantes das pesquisas, a comunidade científica e a sociedade em geral (Academia Brasileira de Ciências, 2013).

A atribuição da qualidade de uma pesquisa ocorre a partir da avaliação dos referenciais teóricos e procedimentos metodológicos adotados, próprios das diferentes áreas de conhecimento e disciplinas. Estas, com frequência, envolvem a obtenção, o tratamento e a análise dos dados que subsidiam os achados e as interpretações do estudo. A atribuição do mérito é feita em termos da originalidade e do impacto dos resultados encontrados.

O *modus operandi* da ciência está em grande parte relacionado ao processo de coleta e análise de dados, componentes que são responsáveis por uma grande parte do tempo e dos recursos utilizados nas pesquisas. Uma parte das disciplinas científicas utiliza-se, em suas pesquisas científicas, de dados coletados no mundo natural, incluindo seres vivos não humanos. Enquanto outra parte, em especial as ciências da saúde e as ciências humanas e sociais, coleta dados diretamente ou relacionados aos seres humanos. Temos ainda que considerar que, do ponto de vista da pesquisa científica, os dados podem ser categorizados como primários, que são aqueles que foram coletados com a finalidade de atender à demanda de um projeto científico, e dados secundários (ou administrativos), aqueles coletados para fins diversos e que eventualmente poderão vir a ser utilizados para a pesquisa científica (Connely et al., 2016).

Quando a pesquisa envolve a coleta de dados diretamente em seres humanos, os imperativos de ética em pesquisa apontam que deve ser autorizada por cada sujeito da pesquisa por meio de um instrumento particular estabele-

cido entre o investigador e o investigado – o Termo de Consentimento Livre e Esclarecido. Essa autorização não está presente no segundo grupo, ou seja, dados pessoais coletados para fins diversos e que eventualmente possam vir a ser utilizados para pesquisa. Essa diferença tem imensa importância no tocante à ética em pesquisa. A inexistência do consentimento do titular dos dados para o uso secundário é um aspecto que tem sido razão de controvérsias envolvendo eticistas e órgãos responsáveis pela regulação ética da pesquisa em seres humanos (Silva et al., 2012).

As ciências que utilizam dados de seres humanos são heterogêneas no tocante à maior ou menor utilização de dados diretamente coletados ou à utilização de dados já coletados. Por exemplo, as ciências da saúde dominantemente utilizam-se de dados que são diretamente coletados pelos seus pesquisadores, enquanto outras, como a economia, privilegiam dados já coletados. A intensificação da digitalização, o aumento da quantidade de dados produzidos e a emergência de novos sistemas de produção e coleta de dados, a exemplo das redes sociais, vêm, de forma rápida e intensa, estimulando e abrindo novas possibilidades para utilização desses dados para a pesquisa. Esse fenômeno, popularmente denominado de *Big Data*, tem sido um dos pilares para a maior centralidade e intensividade no uso de dados observado em muitas ciências, com forte apoio de algoritmos e modelagens computacionais (Leonelli, 2016; Blazquez e Domenech, 2018).

Para responder aos desafios em torno da procedência, do tratamento, da interpretação, da proteção de direitos e do compartilhamento de dados pessoais gerados por projetos de pesquisa, as agências de suporte da pesquisa têm se apoiado na curadoria digital de dados (Digital Curation Centre, 2019) e em princípios internacionalmente aceitos, notadamente os princípios FAIR[4] (Reserach Data Alliance, 2019; *Fostering FAIR Data Principles in Europe*, 2019). Esses princípios delineiam características, ferramentas, vocabulários e infraestruturas para descoberta e reutilização de dados por terceiros (GO FAIR Principles, 2019). Também fazem a distinção entre dados e metadados, para apoiar uma ampla gama de circunstâncias especiais que envolve dados pessoais e sensíveis, visando à conformidade ética e legal.

No entanto, não existem diretrizes de ética em pesquisa harmonizadas em torno da utilização para pesquisa de dados provenientes de fontes externas à comunidade científica. Fazem-se particularmente necessárias reflexões acerca da privacidade de tais dados e das questões éticas e legais que possam ser articuladas ao estabelecimento de políticas, diretrizes, papéis e responsabilidades

4 Os princípios FAIR significam que os dados devem ser encontráveis (*findable*), accessíveis (*accessible*), interoperáveis (*interoperable*) e reutilizáveis (*reusable*).

em torno da gestão de dados e que considere as especificidades da pesquisa científica na era do *Big Data* e dos sistemas informatizados.

Enquanto a identificação da pessoa à qual os dados se referem é central para a definição do que seja dado pessoal, a maioria dos programas científicos prescinde desse tipo de categorização no seu processo de exploração dos dados ou testes de hipóteses. O objetivo de tais programas científicos está no encontro de padrões ou associações sobre grupos de pessoas, emanadas do conjunto de dados analisados. Os resultados são divulgados de maneira agregada, em forma de tabelas, gráficos, sem necessitar fazer referência a qualquer indivíduo em particular. Nessa linhagem de pesquisa, os aspectos éticos a serem escrutinados referem-se à possibilidade de gerar danos ou discriminação, que poderão afetar não propriamente indivíduos particulares, porém grupos de pessoas, como resultado de informações relativas à saúde, *status* socioeconômico, etnia ou outros aspectos contidos nos conjuntos de dados e com potencial de gerar discriminações com relação a esses grupos. Entretanto, mesmo nessa linhagem de pesquisa existe uma etapa em que a identificação pessoal é imprescindível, qual seja, para atender aos objetivos da pesquisa existe a necessidade de integração de diferentes bases de dados. Para essa etapa os riscos necessitam ser avaliados e as ações feitas para mitigá-los.

A INTEGRAÇÃO DE DADOS PARA PESQUISA

No mundo real, os dados pessoais são, em geral, muito diversos e coletados por diferentes agentes e, dessa forma, são acumulados independentes. O uso isolado de bases de dados tem sido uma forma comum de utilização de dados pessoais de origem administrativa ou de outros registros, como os do sistema de saúde. Esses dados são comumente utilizados na pesquisa de-identificados ou agregados em unidades administrativas (p. ex., municípios). Em ambas as situações perdem a condição de dados pessoais.

Porém, a integração (*linkage*) de dados pessoais contidos em diferentes bases tem sido uma abordagem utilizada com crescente frequência, na medida em que, ao agregar um maior número de características dos indivíduos, permite-se que sejam exploradas questões científicas mais complexas.

A pesquisa em saúde reconhece desde há muito tempo a importância da disponibilidade de dados pessoais, seja para a análise clínica de um determinado indivíduo, seja para a pesquisa quantitativa em saúde – é um exemplo o emblemático artigo "*Record linkage*", datado de 1946, que inicia por reconhecer que "*each person in the world creates a Book of Life. This Book starts with birth and ends with death. Its pages are made up of the records of the principal events in life. Record linkage is the name given to the process of assembling the pages of this book into a volume*" (Dunn, 1946).

O processo de integração de dados pessoais pode ocorrer, basicamente, de duas maneiras. Na primeira, quando existe uma identificação numérica única para cada indivíduo, situação que existe em alguns países e torna possível integrar as informações dos indivíduos existentes em diferentes bases de dados pelo uso desse número identificador (integração determinística). No segundo caso, quando não existe esse número identificador, a integração é possível pelo uso de identificadores comuns existentes nas bases a serem integradas, tais como: nome, idade, data e local do nascimento, nome da mãe, etc. (integração probabilística).

Enquanto a vinculação determinística é bastante simples do ponto de vista operacional, a vinculação probabilística exige esforços muito maiores. No caso de bases de dados de grandes populações, ela somente pode ocorrer com a disponibilidade de *softwares* e recursos computacionais adequados. Após a vinculação entre as bases de dados, a base resultante poderá ser de-identificada e transferida para o investigador. Um processo eventualmente utilizado para integração de dados é a pseudoanonimização, definida na Lei Geral de Proteção de Dados (LGPD) como "o tratamento por meio do qual um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo, senão através do uso de informação adicional mantida separadamente pelo responsável em ambiente controlado e seguro".

A vinculação de dados pessoais originários de registros administrativos para estudos de base populacional é uma ferramenta valiosa, por combinar, a nível individual, dados diversificados e provenientes de diferentes fontes. Embora nem sempre substitua os estudos clássicos baseados na coleta de dados primários, as análises dos dados pessoais vinculados vêm tendo o seu uso ampliado em diferentes ciências por apresentar diversas vantagens. Entre outras, mencionamos: são capazes de responder a questões científicas complexas, questões que exigem grandes tamanhos de amostra ou questões relacionadas a populações de difícil acesso. Os estudos, assim gerados, produzem evidências com alto nível de validade externa e, portanto, com maior aplicabilidade para elaboração de políticas públicas.

Existem desafios únicos no uso de dados pessoais com origem em registros administrativos vinculados para pesquisa, entre eles assegurar a confidencialidade do processo. Produzir conjuntos de dados completamente anônimos (onde é possível identificar qualquer indivíduo) seria um elemento de proteção da confidencialidade. No entanto, é cada vez mais claro que o anonimato completo dos dados em nível individual é praticamente impossível, ao mesmo tempo em que seja mantida granularidade suficiente para a pesquisa. Os riscos aumentam dado que tem sido desenvolvidos algoritmos cada vez mais capazes de reidentificar indivíduos em bases de-identificadas (Rocher et al., 2019). Assim, as alternativas para preservação da privacidade durante o processo de in-

tegração de bases de dados devem se pautar não em uma determinada técnica, porém pela combinação de vários procedimentos (Harron et al., 2017):

> (i) Processos claros de acesso a dados pessoais. Isto inclui a existência de base legal, medidas de segurança apropriadas, uso dos dados apenas para finalidade especificada, as credenciais da instituição solicitante, adequada aprovação ética do estudo;
>
> (ii) Definir requisitos do pesquisador, incluindo treinamento e sanções. Os pesquisadores têm a responsabilidade, geralmente definida nos termos de uso, de usar os dados apenas para fins *bona fide*; devem receber treinamento regular em governança da dados; devem existir sanções legais quando os dados são usados de forma inadequada ou sem o devido cuidado;
>
> (iii) Locais físicas ou virtuais estabelecidos para o processamento e vinculação de dados pessoais ou potencialmente identificáveis, que restringem a possibilidade de re-identificação de indivíduos ou mal-uso indevido ou deliberado dos dados. Estes locais são caracterizados por: acordos estritos de acesso, processos seguros de transferência de dados, rede restrita e/ou impossibilidade de acesso à Internet, procedimentos rigorosos de controle de divulgação dos resultados.

Adicionalmente, para que o resultado da vinculação das bases de dados não se torne uma "caixa preta", é necessário que essas sejam associadas a metadados contendo a descrição da proveniência dos dados originais, tratamento aplicado e qualidade das vinculações obtidas para que os pesquisadores julguem a confiabilidade e adequação dos dados aos seus propósitos.

DADOS PARA PESQUISA: DEBATES EPISTEMOLÓGICAS

Como a análise e a interpretação dos resultados oriundos dos dados se baseiam nas questões formuladas mediante a aplicação do método científico e nas técnicas próprias das diferentes áreas de conhecimento e disciplinas, a epistemologia, os desafios à privacidade e a ética entrelaçam-se intimamente nesse novo modo de operar a ciência, apoiada no uso intensivo de grande volume de dados (*Big Data*) (Lipworthy et al., 2017). Desde pelo menos a metade do século XX, o método científico tem sido dominado pela ideia de que o avanço do conhecimento científico ocorre pela elaboração e pelo teste de hipóteses (ou *hypothesis-driven*). Essa tem sido a estratégia mais frequentemente utilizada por cientistas e validada pelas agências de fomento que dão suporte à ciência. Esse modelo implica que o cientista criará ou buscará acesso a um conjunto de dados que o ajude a testar suas hipóteses. Nesse modelo, é central a busca de explicações, muitas vezes causais, dos fenômenos estudados, explicações essas que, em geral, explicam fenômenos ocorridos a nível de grupos populacionais e, portanto, sem pretender abordar o efeito do fenômeno em indivíduos particulares. Esse modelo continua a beneficiar-se do crescimento da disponibilidade de dados, no qual as mesmas hipóteses poderão ser testadas; porém, agora, sem

limitações dos tamanhos amostrais e com maiores possibilidades de generalização dos resultados.

No novo modelo que emerge, despojado de hipóteses previas e que tem sido denominado de *data-driven*, encontra-se como princípio básico a "mineração de dados". Ao passo que seja derivada da crescente disponibilidade de dados, difere do modelo anterior pelo fato de que já não busca mais as explicações dos fenômenos em si, e sim privilegia o desenvolvimento e a sofisticação de algoritmos preditivos. Apesar da crescente demonstração de sua utilidade e a sua aplicação em diferentes campos da ciência como ferramenta para resolver problemas operacionais, muitos tem alertado para existência de riscos relacionados à privacidade, éticos e mesmo políticos, devido ao uso descontrolado dessa abordagem (Mazzocchi, 2015). Com a grande oferta de dados na Internet, em especial dados provenientes de sistemas, aplicativos e plataformas de mídia e redes sociais que captam dados cada vez mais refinados, relacionados aos comportamentos humanos, quando associados a algoritmos cada vez mais eficientes na sua capacidade preditiva, aumenta a capacidade de prever, agora não mais somente comportamentos genéricos de grupos, e sim comportamentos de indivíduos específicos, com todos os seus riscos e dilemas (Zuboff, 2019). Esses algoritmos são a base do moderno *marketing* comercial, que, ao ter maiores conhecimentos sobre o comportamento de indivíduos específicos, pode prever e direcionar as suas necessidades. Porém, essa capacidade tem mostrado que pode ir além, manipulando emoções e comportamentos para alcance de determinados fins, como demonstrado no recente episódio da "Cambridge Analytics" (Andrade, 2018).

Esse famoso caso teve o seu início em uma pesquisa acadêmica de prestigiosa universidade, a qual utilizou o Facebook para recrutar participantes para uma investigação focada na tipificação de perfis de personalidade. Os voluntários baixavam o aplicativo a partir da rede social, e esta estabelecia relações entre os dados do Facebook e de outras ferramentas *on-line* visando ao estabelecimento de métricas para composição do perfil de cada indivíduo que participava do experimento, a partir de suas preferências pessoais. A rede social viabilizou acesso aos dados, sem consentimento, das respectivas redes de amigos dos respondentes da pesquisa, alcançando cerca de 87 milhões de pessoa, cujos dados passaram a ser manipulados com finalidade política em distintos países, sendo os casos mais conhecidos o uso em uma eleição presidencial nos Estado Unidos da América e em um plebiscito no Reino Unido, que buscava decidir a saída ou continuidade do país na Comunidade Europeia (Isaak e Hanna, 2018). Enquanto este episódio tenha ganho grande visibilidade pública, possivelmente este processo de manipulação em comportamentos políticos de indivíduos, próximo ao usado no campo comercial em que os dados são usados para personalizar experiências e otimizar vendas, seja universal e continua a ser

utilizado. Essa nova linha de utilização de dados pessoais, associados a avançados algoritmos preditivos para fins políticos, começa a criar sérias preocupações pelo poder de afetar o equilíbrio até mesmo de democracias estabelecidas. Não por menos que crescem as demandas por formas de regulação para maior proteção e defesa dos direitos das pessoas sobre seus dados e que evitem a utilização para fins não desejados.

PROTEGENDO A PRIVACIDADE DOS SUJEITOS NA PESQUISA CIENTÍFICA: ASPECTOS ÉTICOS

Ao fim da 2ª Guerra Mundial estabeleceu-se a necessidade de serem consolidados princípios éticos universais que protegessem os seres humanos quando participando como sujeito de pesquisas científicas. A inspiração nasce do esforço de renegar os experimentos realizados na Alemanha nazista utilizando seres humanos. Os nazistas utilizaram judeus prisioneiros em campos de concentração para experimentos, nos quais a linha de separação entre tortura e interesse científico era muito débil. Marco importante nesse processo, cuja importância continua na atualidade, é a Declaração de Helsinque[5], de 1964, da Associação Médica Internacional. Essa Declaração estabeleceu princípios e procedimentos fundamentais e universais para a garantia da proteção e dignidade dos sujeitos da pesquisa.

No Brasil, toda pesquisa envolvendo seres humanos requer aprovação do Sistema CEP/Conep, criado em 1996 com o objetivo de proteger o participante da pesquisa e assegurar que o estudo será realizado de acordo com princípios éticos a partir de resoluções e normativas deliberadas pelo Conselho Nacional de Saúde. O Comitê de Ética em Pesquisa (CEP) é a instância institucional e local e a Comissão Nacional de Ética em Pesquisa (Conep) é a instância nacional. A Conep é uma instância colegiada cuja gestão e cujo funcionamento são de responsabilidade do Conselho Nacional de Saúde e da Secretaria de Ciência, Tecnologia e Insumos Estratégicos, vinculada ao Ministério da Saúde.

A avaliação ética do Sistema CEP/Conep é regida por um conjunto de resoluções, destacando-se a Resolução CNS nº 466/2012, que trata de pesquisas envolvendo seres humanos[6] e normas complementares, a exemplo da Resolução CNS nº 580/2018, voltada a especificidades éticas das pesquisas de interesse estratégico para o Sistema Único de Saúde (SUS)[7], e da Resolução

5 Disponível em: <https://www.wma.net/wp-content/uploads/2016/11/491535001395167888_DoHBrazilianPortugueseVersionRev.pdf>.

6 BRASIL. Conselho Nacional de Saúde. Resolução nº 466, de 12 de dezembro de 2012. Disponível em: <http://conselho.saude.gov.br/resolucoes/2012/Reso466.pdf>.

7 BRASIL. Conselho Nacional de Saúde. Resolução nº 580/2018. Disponível em: <https://conselho.saude.gov.br/resolucoes/2018/Reso580.pdf>.

nº 510/2016, voltada a especificidades éticas das pesquisas que utilizam metodologias próprias das Ciências Humanas e Sociais[8].

O Sistema CEP/Conep reconhece o Termo de Consentimento e Assentimento Livre e Esclarecido do participante ou de seu responsável legal como o principal instrumento para garantir a ética de um projeto de pesquisa. Os termos visam demonstrar autonomia e liberdade de participação a partir de esclarecimentos sobre a natureza da pesquisa, objetivos, métodos, benefícios previstos, potenciais riscos e incômodos que a participação no estudo poderão acarretar, facultando a participação voluntária, a requisição de informações adicionais e a retirada do consentimento a qualquer tempo sem prejuízo algum ao participante. Além da manifestação de consentimento do participante, a avaliação ética de uma pesquisa baseia-se em um conjunto de fundamentos, entre os quais estão a relevância social da pesquisa e o compromisso com o máximo de benefícios e o mínimo de danos e riscos.

Na impossibilidade de obtenção de termo de consentimento, a dispensa poderá ser solicitada pelo pesquisador ao Sistema CEP/Conep, desde que seja justificada. O pedido de dispensa deve ser fundamentado e avaliado em termos de responsabilidades assumidas pelos pesquisadores para mitigar riscos e assegurar os direitos dos participantes, incluindo procedimentos voltados à confidencialidade e à privacidade dos titulares dos dados, à proteção da imagem e à não estigmatização de indivíduos ou grupos.

No caso de dados pessoais que não são primariamente coletados para pesquisa e que não contam com TCLE, o uso secundário em pesquisa tem sido intensamente feito em diferentes sociedades e, enquanto não sem algumas controvérsias (Breen 2001; Silva et al., 2012), de um modo geral, os organismos de regulação ética têm aprovado o seu uso.

PROTEGENDO A PRIVACIDADE DOS SUJEITOS NA PESQUISA CIENTÍFICA: ASPECTOS LEGAIS

Nas recentes legislações sobre proteção de dados pessoais da União Europeia[9] e do Brasil[10], a pesquisa científica é considerada uma hipótese legítima para o tratamento secundário de dados pessoais, devendo ser observados o respeito aos padrões éticos relevantes da área de conhecimento, o interesse legítimo do uso (relevância e benefício), a avaliação das necessidades e a proporcionalidade das operações de tratamento de dados pessoais em relação à finalidade (princípio da minimização de danos), bem como a avaliação de riscos

8 BRASIL. Conselho Nacional de Saúde. Resolução nº 510/2016. Disponível em: <http://bvsms.saude.gov.br/bvs/saudelegis/cns/2016/res0510_07_04_2016.html>.

9 Disponível em: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679>.

10 BRASIL. Lei nº 13.709, de 14 de agosto de 2018.

aos direitos e às liberdades das pessoas e as medidas previstas para controlar os riscos em termos de acesso indevido e vazamento de dados.

Até 2018, o Brasil possuía um quadro regulatório tímido e insuficiente sobre o tratamento de dados pessoais no território nacional. A despeito da Lei de Acesso à Informação (LAI) regulamentada pelo Decreto nº 7.724/2012 e de outras leis setoriais, não existia normativa específica que assegurasse conformidade jurídica em torno do tratamento de dados pessoais no País (Guanaes et al., 2018). O marco da regulamentação da proteção de dados pessoais no País é a Lei Geral de Proteção de Dados Pessoais (LGPD), sancionada e publicada em agosto de 2018 e com vigência prevista a partir de agosto de 2020. A LGPD se aplica a qualquer operação de tratamento de dados pessoais realizada por pessoa natural ou por pessoa jurídica de direito público ou privado no território nacional.

A lei prevê que dados pessoais e sensíveis devam ser tratados de forma legal, justa e transparente em relação aos titulares dos dados, em decorrência dos potenciais riscos em relação a seus direitos e suas liberdades. Portanto, a conformidade com a lei exige que os responsáveis pelo tratamento dos dados pessoais organizem e mantenham registros claros e seguros acerca de qualquer atividade relacionada ao processamento desses dados sob sua responsabilidade, pois os titulares devem ter acesso facilitado às informações sobre qualquer tratamento pelos quais seus dados passem.

A Lei Geral de Proteção de Dados Pessoais brasileira foi inspirada, em parte, pelo Regulamento Geral de Proteção de Dados da União Europeia (GDPR). No seu art. 4º, a lei brasileira estabelece seu escopo e define que não se aplica ao tratamento de dados pessoais realizado para fins exclusivamente jornalístico, artístico e, também, acadêmico. Para este último, remete para os arts. 7º e 11. O art. 7º define as condições para uso dos dados pessoais entre as quais reza "para a realização de estudos por órgão de pesquisa", com a ressalva de que deve ser "garantida, sempre que possível, a anonimização". O art. 11 veda o uso de dados sensíveis (conforme definido na lei, são "dados pessoais sobre a origem racial ou étnica, as convicções religiosas, as opiniões políticas, a filiação a sindicatos ou a organizações de caráter religioso, filosófico ou político, dados referentes à saúde ou à vida sexual, dados genéticos ou biométricos, quando vinculados a uma pessoa natural"), sendo possível o seu uso apenas com o consentimento específico do titular. Porém, também estabelece algumas exceções, entre as quais para a "realização de estudos por órgão de pesquisa, sendo garantida, sempre que possível, a anonimização". Por fim, o art. 13 foca na pesquisa em saúde pública e estabelece que

> os órgãos de pesquisa poderão ter acesso a bases de dados pessoais, que serão tratados exclusivamente dentro do órgão e estritamente para a finalidade de realização de estudos e pesquisas e mantidos em ambiente controlado e seguro, con-

> forme práticas de segurança previstas em regulamento específico e que incluam, sempre que possível, a anonimização ou pseudoanonimização dos dados, bem como considerem os devidos padrões éticos relacionados a estudos e pesquisas.

Entendemos que, para a lei, as atividades de pesquisa são consideradas enquanto um contexto específico de processamento de dados pessoais, que deve equilibrar os direitos individuais e a busca pelo interesse público a partir da aplicação de medidas técnicas e organizacionais suficientes e adequadas para garantir a proteção dos dados e o mínimo possível de processamento, possibilitando que sejam alcançados os objetivos das pesquisas, reduzindo os riscos relacionados a sua utilização. A despeito de prever a proibição geral do processamento de dados pessoais sensíveis, existem exceções para prática de cuidados de saúde, saúde pública e alguns setores de pesquisa em que o tratamento é autorizado sob condições específicas.

Destaca-se que a lei brasileira aborda o tratamento de dados para pesquisa em termos específicos, pois se refere exclusivamente a órgãos de pesquisa definidos na própria LGPD como

> órgão ou entidade da Administração Pública direta ou indireta ou pessoa jurídica de direito privado sem fins lucrativos legalmente constituída sob as leis brasileiras, com sede e foro no País, que inclua em sua missão institucional ou em seu objetivo social ou estatutário a pesquisa básica ou aplicada de caráter histórico, científico, tecnológico ou estatístico. (Art. 5º, inciso XVIII)

Além de ambiente controlado e seguro, a anonimização e a pseudoanonimização dos dados são tidas como as principais técnicas que deverão ser adotadas no tratamento dos dados pessoais para proteção da privacidade dos indivíduos.

A lei define a anonimização como a "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" (art. 5º, inciso XI). A lei não se aplica aos dados anonimizados. Caso a anonimização seja revertida, os dados passam a ser considerados dados pessoais e as disposições da lei são aplicadas. Entende-se que o risco de reversão da anonimização se relaciona ao interesse e aos esforços de reidentificação das pessoas, que se configuram em contravenção.

Por sua vez, o conceito de pseudoanonimização é definido na lei como "o tratamento por meio do qual um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo, senão pelo uso de informação adicional mantida separadamente pelo controlador em ambiente controlado e seguro" (art.13, inciso IV). Na lei, os dados pseudoanonimizados são considerados dados pessoais pela possibilidade de rastrear os dados de volta ao indivíduo por meio do código-chave.

OS CENTRO DE DADOS ORIENTADOS PARA PESQUISA

Na LGPD, as instituições de pesquisa são responsáveis por aplicação e zelo dos preceitos legais. Pela necessidade de infraestrutura adequada, pessoal especializado e governança de dados, chamamos atenção para existência de centros de dados pessoais criados em alguns países para prover o acesso a dados de qualidade, de forma segura e controlada para pesquisa, avaliação, planejamento e elaboração de políticas.

Verifica-se que, em geral, os centros de dados são criados por meio de parcerias entre governos, universidades e instituições de pesquisa, a exemplo do Manitoba Centre for Health Policy, no Canadá[11]; Massive Data Institute, nos Estados Unidos[12]; Centre for Big Data Research in Health, na Australia[13]; Administrative Data Research Centres, no Reino Unido[14]; Integrated Data Infrastructure, na Nova Zelândia[15]; e o Sail Databank, no País de Gales, considerado a maior e mais acessível fonte de dados vinculados e anonimizados para pesquisa em escala populacional do mundo[16]. No Brasil, temos o exemplo do Centro de Integração de Dados e Conhecimentos para a Saúde (CIDACS), que é uma iniciativa da Fiocruz em colaboração com pesquisadores de diversas Universidades (Barreto et al., 2019) e que tem como sua missão central processar grandes bases de dados nacionais da área da saúde e de áreas relacionadas para, ao integrar tais bases, proporcionar *datasets* desidentificados que possam ser usados para a pesquisa e a avaliação.

Esses centros agregam grandes quantidades de dados pessoais, que têm origem em registros administrativos. Eles se originam em departamentos ou agências governamentais para prestação de serviços ou administração de programas governamentais, como, por exemplo, educação, programas de proteção social, habitação, censo populacional, etc., ou outros registros eletrônicos, como os registros de saúde. Esses dados compreendem informações sobre indivíduos que não foram coletadas para fins de pesquisa, mas que podem ser preparadas e integradas para uso em projetos de pesquisa.

O estabelecimento de centros de dados a partir de parcerias entre o governo e os órgãos de pesquisa, além de otimizar recursos e garantir sustentabilidade, viabiliza a colaboração entre agências governamentais no que concerne

11 Manitoba Centre. About MCHP. Disponível em: <http://umanitoba.ca/faculties/health_sciences/medicine/units/chs/departmental_units/mchp/about.html>.

12 Massive Data Institute. About MDI. Disponível em: <https://mccourt.georgetown.edu/massive-data-institute/about>.

13 Centre for Big Data Research in Health. Disponível em: <https://cbdrh.med.unsw.edu.au/>.

14 Administrative Data Research Centres no Reino Unido. Disponível em: <https://www.ons.gov.uk/aboutus/whatwedo/programmesandprojects/theadministrativedataresearchnetworkcollaboration>.

15 Integrated Data Infrastructure. Disponível em: <https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/>.

16 Disponível em: <https://saildatabank.com/about-us/overview/>.

ao compartilhamento de dados, documentação associada, estabelecimento de padrões de coleta, tratamento, armazenamento e preservação de dados que têm implicações diretas sobre a governança e qualidade de dados utilizados em pesquisas de base populacional. Centros de dados dessa natureza viabilizam acesso a dados integrados e desidentificados para pesquisa a partir de arranjos de segurança apropriados e estabelecimento de requisitos para aprovação do usuário. Em geral, os resultados são disponibilizados aos pesquisadores após cuidadosa de-identificação com a retirada do máximo possível de potencias identificadores, no sentido de evitar a reidentificação dos indivíduos[17].

A partir de termos e condições estabelecidos pelos centros de dados, que incluem orientações acerca do acesso e da utilização dos dados, os pesquisadores assumem a responsabilidade de usar os dados apenas para fins legítimos, com compromisso de uma prática científica *bona fide*, bem como estar cientes de que ações legais serão tomadas se os dados forem utilizados inadequadamente ou sem o devido cuidado.

Além da necessidade de seguir preceitos éticos e legais, a proteção da privacidade dos indivíduos por meio de centros de dados é estratégica para boas práticas de gestão por parte do Estado e da comunidade científica, pois aumentam a confiança da sociedade na utilização de seus dados para obtenção de conhecimentos e evidência que visem responder perguntas e solucionar problemas de interesse da sociedade. Por exemplo, no Reino Unido, com o estabelecimento da Administrative Data Research Network, criou-se a possibilidade de utilizar dados administrativos como alternativa ao próximo Censo, que ocorrerá em 2021. O Conselho de Pesquisa Econômica e Social (ESRC) e o Escritório de Estatísticas Nacionais (ONS) daquele país encomendaram estudo com o objetivo de explorar questões em torno da visão da sociedade sobre o uso de dados administrativos contendo informações pessoais para pesquisa. O estudo aponta que a sociedade tende a aprovar o uso de conjuntos de dados administrativos para pesquisa que apresente potencial benefício para sociedade, desde que os dados sejam fornecidos desidentificados e que sejam tratados, acessados e mantidos em local adequado e seguro (Cameron et al., 2014).

Além da governança de dados que envolve medidas técnicas e administrativas para o provimento adequado de dados em termos éticos e legais, chamamos atenção para importância do rigor metodológico em torno do tratamento e da vinculação de grande volume de dados administrativos.

17 Disponível em: <https://www.adruk.org/our-mission/ethics-responsibility/>.

CONSIDERAÇÕES FINAIS

A LGPD é uma lei geral voltada ao estabelecimento de princípios e conceitos norteadores para preservar o equilíbrio entre a necessidade de proteger efetivamente os direitos dos titulares dos dados, ao mesmo tempo em que permite o processamento de dados pessoais e sensíveis para fins determinados, inclusive a pesquisa científica. O respeito a padrões éticos é parte da legalidade do processamento de dados pessoais e sensíveis em pesquisa, que deverá ser consistente com normatização específica do setor, que, no caso, será de responsabilidade do Sistema CEP/Conep e da Autoridade Nacional de Proteção de Dados em diálogo com a comunidade científica para definir o que necessitará ser regulado e normatizado como desdobramento da LGPD.

Levando-se em consideração que a LGPD é recente, e que o País não possui experiência prévia em termos de legislação voltada a proteção de dados pessoais, acreditamos ser necessário aprender com experiências mais maduras de outros países para adequá-las a nossa realidade e para buscar a interoperabilidade legal no nível internacional para que a lei brasileira seja protetiva e não inviabilizadora da pesquisa científica.

Dados administrativos não são coletados com finalidade de pesquisa, e a sua transformação em fonte de informações apresenta um conjunto de desafios metodológicos relacionados à privacidade, à ética, à regulação do acesso, ao pré-processamento das bases originais (seleção, limpeza, padronização e harmonização das variáveis) e à utilização de algoritmos adequados aos tipos e tamanhos das bases de dados para serem adequadamente vinculados. Todo esse processo feito com o objetivo de contribuir com conhecimentos para a sociedade, garantindo alto níveis de privacidade e ética.

REFERÊNCIAS

ABC. Academia Brasileira de Ciências. Rigor e Integridade na Condução da Pesquisa Científica. Guia de Recomendação de Práticas Responsáveis. Rio de Janeiro: Academia Brasileira de Ciências, 2013. Disponível: <http://www.abc.org.br/IMG/pdf/doc-4311.pdf>. Acesso em: 25 ago. 2019.

ANDRADE, Diogo Queiroz de. Facebook. Cambridge Analytica, a empresa que manipula a democracia à escala global. Público, mar. 2018. Disponível em: <https://www.publico.pt/2018/03/20/tecnologia/noticia/ca-a-empresa-que-manipula-a-democracia-a-escala-global-1807409>.

BARRETO, M. L.; ICHIHARA, M. Y.; ALMEIDA, B. A.; BARRETO, M. E.; CABRAL, L.; FIACCONE, R. L.; CARREIRO, R. P.; TELES, C. A.; PITTA, R.; PENNA, G. O.; BARRAL-NETTO, B.; ALI, M. S.; DENAXAS, S.; RODRIGUES, L. C.; SMEETH, L. The Centre for Data and Knowledge Integration for Health (CIDACS): Linking Health and Social Data in Brazil. *International Journal of Population Data Science, in press*.

BLAZQUEZ, D.; DOMENECH, J. Big Data sources and methods for social and economic analyses. *Technological Forecasting & Social Change*, v. 130, p. 99-113, May 2018.

BREEN, K. J. Consent for the linkage of data for public health research: is it (or should it be) an absolute pre-requisite? *Aust N Z J Public Health*, 25(5):423-5, 2001.

CAMERON, D.; POPE, S.; CLEMENCE, M. Dialogue on Data: Exploring the public's views on using administrative data for research purposes. Office for National Statistics; Economic & Social Research Council, 2014. Disponível em: <https://esrc.ukri.org/files/public-engagement/public-dialogues/dialogue-on-data-exploring-the-public-s-views-on--using-linked-administrative-data-for-research-purposes/>.

CONNELLY, R.; PLAYFORD, C. J.; GAYLE, V.; DIBBEN, C. The role of administrative data in the big data revolution in social science research. *Social Science Research*, 59, p. 1-12, 2016.

DCC. Digital Curation Centre. What is digital curation? Disponível em: <http://www.dcc.ac.uk/digital-curation/what-digital-curation>. Acesso em: 19 ago. 2019.

DUNN, Halbert. Record linkage. *American Journal of Public Health*, p. 1412-1416, 1946.

FOSTERING FAIR Data Practices in Europe. Disponível em: <https://fairsfair.eu/news/basics-eosc-and-fair>. Acesso em: 25 ago. 2019.

GO FAIR. FAIR Principles, 2019. Disponível em: <https://www.go-fair.org/fair-principles/>. Acesso em: 25 ago. 2019.

GUANAES, P.; SOUZA, Allan Rocha de; DONEDA, Danilo; NASCIMENTO, Francisco José Tavares do. Marcos legais nacionais em face da abertura de dados para pesquisa em saúde. Dados pessoais, sensíveis ou sigilosos e propriedade intelectual/Paulo Guanaes; Allan Rocha de Souza; Danilo Doneda; Francisco José Tavares do Nascimento. Rio de Janeiro: Fiocruz, 2018. 122 p. Disponível em: <https://www.arca.fiocruz.br/bitstream/icict/28838/4/Guanaes_Paulo_Org_Marcos_Legais_Presid%C3%AAncia_2018.pdf.>

HARRON, K.; DIBBEN, D.; BOYD, J.; HJERN, A.; AZIMAEE, M.; BARRETO, M.; GOLDSTEIN, H. Challenges in administrative data linkage for research. *Big Data & Society*, 4(2):1-12, July/December 2017.

ISAAK, J.; HANNA, M. J. User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. The Policy Corner, ago. 2018. Disponível em: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8436400&tag=1>.

KUHN, T. S. *A estrutura das revoluções científicas*. 9. ed. (2006). São Paulo: Perspectiva, 2009.

LEONELLI, S. *Data-Centric Biology*: A Philosophical Study. The University Chicago Press: Chicago and London, 2016.

LIPWORTH, W.; MASON, P. H.; KERRIDGE, I. et al. Ethics and Epistemology in Big Data Research. *Bioethical Inquiry*, 14: 489-500, 2017.

MAZZOCCHI F. Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science. *EMBO Rep.*, 16(10):1250-1255, 2015.

MERTON, R. K. A ciência e a estrutura social democrática. In: *Ensaios de sociologia da ciência*. 1. ed. São Paulo: Associação Filosófica Scientiae Studia/Editora 34, 2013. p. 181-198.

RESEARCH Data Alliance. FAIR. Disponível em: <https://www.rd-alliance.org/fair>. Acesso em: 25 ago. 2019.

ROCHER, L.; HENDRICKX, J. M.; DE MONTJOYE, Y. A. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun.*, 10(1):3069, 2019.

SILVA, M. E.; COELI, C. M.; VENTURA, M.; PALACIOS, M.; MAGNANINI, M. M.; CAMARGO, T. M.; CAMARGO JR., K. R. Informed consent for record linkage: a systematic review. *J Med Ethics*, 38(10):639-42, Oct. 2012.

ZUBOFF, S. *The age of surveillance capitalism*: the fight for a human future at the new frontier of power. New York: Public Affairs, 2018.

Data de submissão: 30.09.2019
Data de aceite: 27.10.2019