

Dados governamentais na perspectiva da Ciência Aberta: potencialidades e desafios para saúde pública a partir de um estudo de caso

Bethânia Almeida
Fundação Oswaldo Cruz
bethania.almeida@fiocruz.br

Paula Xavier
Fundação Oswaldo Cruz
paula.xavier@fiocruz.br

Mauricio Barreto
Fundação Oswaldo Cruz
mauricio.barreto@fiocruz.br

Resumo

O acesso, uso e reutilização de dados governamentais administrativos com finalidade de pesquisa, principalmente aqueles contendo informações pessoais, são tópicos amplamente discutidos atualmente. Levando-se em consideração que muitos países estão em fase incipiente de adoção de políticas de acesso a esses dados, acreditamos que a experiência do Centro de Integração de Dados e Conhecimento para Saúde (CIDACS) da Fiocruz poderá aportar contribuições ao debate, pois questões relativas ao acesso, uso e reutilização de dados administrativos estão sendo discutidas e executadas no CIDACS, inclusive para aderir a princípios da ciência aberta.

Palavras chave: Dados Governamentais; Integração de Dados; Saúde Pública; Ciência Aberta.

Government data from the perspective of Open Science: potential and challenges for public health

Abstract

Access to, use of, sharing and reuse of governmental administrative datasets for research purposes, principally ones containing personally identifiable information, are all topics being widely discussed nowadays. In fact, as many countries are in the

incipient stages of adopting access policies, we believe that our experience at the Fiocruz Center for Data and Knowledge Integration in Health (CIDACS) can contribute significantly. Relevant issues pertaining to the access, use and reuse of administrative data are currently being discussed and implemented at CIDACS, which may eventually become adherent to the principles of open science.

Key words: Governmental Data; Data Integration; Public health; Open Science.

Introdução

Atualmente, a Ciência Aberta é amplamente defendida em termos de transparência, reprodutibilidade, disseminação e transferência de novos conhecimentos. Devido à complexidade do tema, notamos que as suas práticas se concentraram principalmente em dois aspectos: acesso aberto a publicações científicas e, mais recentemente, abertura de dados.

Um evento decisivo no esforço para fornecer acesso aberto ao conhecimento científico foi a Budapest Open Access Initiative / BOAI, publicada em 2002, que propôs conceitos e estratégias para alcançar o acesso aberto através de periódicos (via dourada) e auto-arquivamento em repositórios públicos (via verde).

Nos últimos anos a abertura de dados científicos tornou-se uma das principais propostas da Ciência Aberta em meio a um intenso debate sobre os princípios e padrões que deverão ser adotados. Um documento publicado em 2007 pela Organização para Cooperação e Desenvolvimento Econômico foi um marco nessa discussão visando fomentar a cooperação internacional para o compartilhamento de dados gerados por pesquisas financiadas com recursos públicos. Sendo as agências de financiamento instrumentais na formulação de políticas para gestão, compartilhamento e reutilização de dados, pois a solicitação de planos de gestão de dados tornou-se parte dos requisitos obrigatórios para financiar projetos de pesquisa (SANTOS et al., 2017). Mais recentemente, revistas científicas também passaram a solicitar dados que subsidiam os achados dos artigos submetidos.

Dados sempre foram ativos importantes para produção de conhecimento, cuja aquisição e tratamento para análise demandam a maior parte do tempo e dos recursos financeiros destinados às pesquisas. Nesta perspectiva, verifica-se certo consenso internacional acerca das recomendações e políticas relacionadas ao acesso, compartilhamento e preservação de dados gerados por projetos de pesquisa, que geralmente ocorrem via repositórios de áreas disciplinares, repositórios institucionais e repositórios públicos.

Em relação aos dados governamentais, um consenso ainda não foi alcançado para abordar questões pertinentes ao acesso, uso e reutilização com finalidade de pesquisa, principalmente aqueles que contêm informações pessoais. Tópicos que têm sido amplamente

discutidos, sendo parte de um esforço para manter a confidencialidade, proteger a privacidade individual, respeitar os termos de consentimento e gerenciar a segurança da informação.

Neste trabalho nos referimos especificamente a dados governamentais administrativos, informações coletadas por qualquer departamento ou agência do governo em seus serviços rotineiros a exemplo de registros escolares e informações sobre saúde (Administrative Data Research Network, 2018). Dados concebidos enquanto potenciais fontes de conhecimento para que pesquisadores, formuladores de políticas, gestores e sociedade em geral decidam como lidar com uma série de questões e desafios complexos em diferentes esferas da vida da sociedade humana.

A despeito da abertura de dados governamentais constar na taxonomia da Ciência Aberta (Foster, 2018) e ser fonte para geração de evidências para formulação de políticas públicas (Open Government Partnership, 2018), o seu acesso é mais voltado à transparência, acesso à informação pública e participação social.

Levando-se em consideração que o acesso a dados governamentais com finalidade de pesquisa é restrito em muitos países tendo, portanto, desafios de ordem ética e legal para aderir aos princípios da Ciência Aberta, o artigo aborda o tema a partir da experiência de um centro de integração de dados voltado à pesquisa em saúde pública.

Saúde Pública e Dados Governamentais

A saúde pública visa promover e proteger a saúde de pessoas e comunidades contemplando em suas abordagens aspectos sociais, econômicos e ambientais que impactam as condições em que as pessoas nascem, crescem, vivem, trabalham e envelhecem (WHO, 2017). Necessitando, portanto, integrar conhecimentos e evidências intra e entre campos disciplinares para compreensão de efeitos isolados ou associados sobre a saúde de pessoas e populações para ampliar as chances de obtenção de respostas (conhecimento) e soluções (políticas públicas).

Nesta perspectiva, o uso de dados administrativos integrados com dados de outras fontes possibilita responder perguntas de pesquisa de forma mais rápida, econômica e confiável quando comparado com métodos tradicionais que subsidiam estudos de base populacional por exemplo. Nesses casos, dados individualizados são necessários para integração de dados dos mesmos indivíduos oriundos de outras fontes suscitando discussões para manter a confidencialidade, proteger a privacidade individual, respeitar os termos de consentimento e gerenciar a segurança da informação.

Em alguns países do mundo esta abordagem de integração de dados tem sido extensivamente utilizada, como a Dinamarca (Davidsen 2011; Schmidt 2014), que organizaram coortes para pesquisa, com dados de toda população. No Reino Unido, a possibilidade de explorar a grande base de dados gerados pelo Sistema Nacional de Saúde é parte das estratégias para dinamizar o processo de inovação nas ciências da vida (UK, 2011).

O acesso a dados governamentais administrativos contendo informações pessoais é restrito em muitos países. Em alguns casos, o acesso é permitido quando existe eminente interesse público a exemplo do novo Regulamento Geral de Proteção de Dados da União Europeia (UE 2016/679), que entrou em vigor a partir de 25 de maio de 2018. O Regulamento possui casos específicos sobre a utilização de dados pessoais para pesquisa, incluindo o processamento de dados que merecem proteção especial para propósitos relacionados à saúde quando houver relevância e benefício para o cidadão e para a sociedade.

No Brasil, a lei geral de proteção e tratamento de dados pessoais foi sancionada com vetos em 14 de agosto de 2018. A lei visa regulamentar a proteção de dados pessoais, tanto pelo poder público quanto pela iniciativa privada. Dentre outros aspectos disciplina as regras para acesso a bases de dados pessoais por órgãos de pesquisa. Na realização de estudos em saúde pública, os órgãos de pesquisa poderão ter acesso a bases de dados pessoais, que serão tratados exclusivamente dentro do órgão e mantidos em ambiente controlado e seguro, incluindo sempre que possível a anonimização ou pseudomização dos dados.

Nas regulamentações mencionadas, os potenciais benefícios de ampliar conhecimentos e gerar evidências que subsidiem decisões regulatórias e de políticas públicas na área da saúde são reconhecidos para o acesso e uso de dados pessoais para a pesquisa, desde que empregadas medidas técnicas e administrativas de segurança para evitar a violação da privacidade dos indivíduos.

Nesta perspectiva, apresentamos o caso do Centro de Integração de Dados e Conhecimentos para a Saúde da Fiocruz, por acreditar que a experiência poderá contribuir com discussões voltadas à superação de desafios para o acesso e uso de dados governamentais administrativos com finalidade de pesquisa na perspectiva da Ciência Aberta.

Estudo de Caso

O Centro de Integração de Dados e Conhecimento para a Saúde (CIDACS) da Fiocruz foi inaugurado em 2016 com a missão de realizar pesquisas, desenvolver metodologias investigativas e promover a capacitação profissional e científica a partir da integração de grande volume de dados oriundos de distintas fontes para ampliar o entendimento dos determinantes sociais e ambientais sobre a saúde da população brasileira, além de apoiar a tomada de decisões em políticas públicas em benefício da sociedade. Para alcançar seus propósitos voltados à integração de dados para apoiar estudos de base populacional, dados governamentais administrativos contendo informações pessoais são essenciais. Dados cedidos mediante solicitação formal e assinatura de termos de responsabilidade.

A partir de sua criação passou a enfrentar vários desafios para desenvolver uma plataforma de dados integrados para pesquisa com segurança da informação, ética e privacidade transversais em todos os processos e níveis visando:

- Receber, tratar e processar grande volume de dados, aspectos que incluem vinculação, estimativa de acurácia, extração e anonimização;
- Armazenar, catalogar e preservar datasets;
- Acessar os dados vinculados e anonimizados;
- Desenvolver métodos para análise.



Figura 1. Arquitetura da plataforma de dados do Cidacs

Fonte: Cidacs, 2018

Depois de receber autorização para usar os bancos de dados administrativos necessários para os projetos de pesquisa propostos, conforme aprovação prévia por um comitê de ética, esses conjuntos de dados são recebidos juntamente com todos os protocolos de segurança da informação e ingeridos na Plataforma. Ambiente projetado para preservar a confidencialidade por meio de uma combinação de configurações físicas e virtuais que restringem a possibilidade de reidentificação de indivíduos de acordo com considerações para ambientes seguros de vinculação de dados, incluindo a separação de processos de vinculação e análise (Harron et al. , 2017).

A preparação e processamento dos dados envolvem áreas de ciência de dados, curadoria digital e estatística visando garantir a qualidade dos dados vinculados e anonimizados. Processos descritos detalhadamente em metadados para permitir a avaliação da qualidade e confiabilidade para uso e reuso em pesquisas.

Os pesquisadores têm permissão apenas para acessar variáveis de dados codificados e não identificados relevantes para suas respectivas pesquisas presencialmente ou via Rede Virtual Privativa com provimento de ferramentas de visualização e análise, mediante assinatura de termos de responsabilidade com condições de uso. No momento, somente os projetos de

pesquisa de membros internos e colaboradores são considerados qualificados para usar os bancos de dados vinculados e anonimizados.

Ainda em 2018 serão publicados metadados com identificadores persistentes acompanhados de termos e condições para acesso aos dados em si. Desse modo, a existência e descrição dos dados se tornam públicas e, quando viável, será fornecido acesso controlado aos dados em questão. Aspectos convergentes com os princípios FAIR, adotados como padrão internacional para abertura de dados na perspectiva da Ciência Aberta.

Os princípios FAIR foram elaborados para permitir que dados fragmentados e dispersos se tornem Encontráveis (Findable), Acessíveis (Accessible), Interoperáveis (Interoperable) e Reutilizáveis (Reusable), delineando as características, ferramentas, vocabulários e infraestruturas necessários para descoberta e reutilização de dados por terceiros (Wilkinson et al., 2016; European Commission, 2016; Go Fair, 2018). Esses princípios distinguem dados e metadados para apoiar uma ampla gama de circunstâncias especiais, incluindo dados pessoais e sensíveis.

A governança de dados no Cidacs leva em consideração bases legais e éticas no acesso e tratamento de dados bem como nos termos e condições de uso dos dados vinculados e anonimizados, que dentre outros aspectos abarca medidas apropriadas de segurança, credenciais adequadas para requisitar os dados e sanções legais quando os dados são utilizados de forma inapropriada ou sem o devido cuidado.

A governança de dados no Cidacs está alinhada com o contexto de vanguarda de vinculação de dados para pesquisa em saúde pública com ética, segurança e privacidade visando garantir direitos individuais e coletivos. Adicionalmente, passou a empreender esforços para viabilizar a reutilização dos dados produzidos em formatos aderentes aos princípios da Ciência Aberta.

Considerações finais

Por sua natureza e missão, que envolvem pesquisa, geração e processos de conhecimento utilizando dados, o Cidacs se interessa por questões relacionadas ao acesso e compartilhamento de dados com finalidade de pesquisa e geração de evidências para políticas públicas, passando a desenvolver estudos para que sua governança de dados esteja alinhada com o contexto de vanguarda no âmbito internacional de vinculação de dados para produção de conhecimento para saúde pública com ética, segurança e privacidade e com aderência aos princípios da Ciência Aberta.

Referências

Administrative Data Research Network, Glossary. (2018). Disponível em: <https://www.adrn.ac.uk/public-engagement/glossary/>.

Budapest Open Access Initiative. Disponível em:

<http://www.budapestopenaccessinitiative.org/read>

Adrian Spoerri, Marcel Zwahlen, Matthias Egger, Matthias Bopp. The Swiss National Cohort: a unique database for national and international researchers. *Int J Public Health* (2010) 55:239–242 DOI 10.1007/s00038-010-0160-5

Davidson M, Kjølner M, Helweg-Larsen K. The Danish National Cohort Study (DANCOS). *Scand J Public Health*. 2011 Jul;39(7 Suppl):131–5. doi: 10.1177/1403494811399167.

European Commission. (2016). H2020 Programme: guidelines on FAIR Data Management in Horizon 2020, v.3. Disponível em:

http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

Foster Open Science. (2018). Open Science Taxonomy, Open Government Data. Disponível em: <https://www.fosteropenscience.eu/foster-taxonomy/open-government-data>.

Go Fair. (2018). Global Open, Findable, Accessible, Interoperable, Reusable. Disponível em: <https://www.go-fair.org/>.

Harron K., Dibben, D., Boyd, J., Hjern, A., Azimae, M., Barreto, M., Goldstein, H. Challenges in administrative data linkage for research. *Big Data & Society*. July–December 2017: 1–12. 017. Disponível em: <http://journals.sagepub.com/doi/pdf/10.1177/2053951717745678>

Khoury, M.J; Ioannidis, J.P.A. Big data meets public health. *Science* Vol. 346, Issue 6213, 1054–1055, 2014.

Lea, N. C., Nicholls, J., Dobbs, C., Sethi, N., Cunningham, J., Ainsworth, J., Kalra, D. (2016). Data Safe Havens and Trust: Toward a Common Understanding of Trusted Research Platforms for Governing Secure and Ethical Health Research. *JMIR Medical Informatics*, 4(2), e22. Disponível em: <http://doi.org/10.2196/medinform.5571>

Open Government Partnership. (2018). About OGP. Disponível em: <https://www.opengovpartnership.org/about/about-ogp>.

Organização para Cooperação e Desenvolvimento Econômico. Principles and Guidelines for Access to Research Data from Public Funding. Disponível em: <http://www.oecd.org/sti/sci-tech/38500813.pdfv>

Santos, P. X, Almeida, B. Henning, P. (org). Livro Verde – Ciência aberta e dados abertos: mapeamento e análise de políticas, infraestruturas e estratégias em perspectiva nacional e internacional. Rio de Janeiro: Fiocruz, 140 p. 2018. Disponível em: <https://www.arca.fiocruz.br/handle/icict/24117>

Schmidt M, Pedersen L, Sørensen HT. The Danish Civil Registration System as a tool in epidemiology. *Eur J Epidemiol*. 2014 Aug;29(8):541–9. doi: 10.1007/s10654-014-9930-3.

Silva, M. E.M. Linkage de bases de dados identificadas em saúde: consentimento, privacidade e segurança da informação. Tese – Universidade Federal do Rio de Janeiro, 2012. Disponível em: <http://www.posgraduacao.iesc.ufrj.br/media/tese/1363187343.pdf>

UK Department for Business, Innovation and Skills and Office for Life Sciences. Strategy for UK Life Sciences, 2011. Disponível em:
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/32457/11-1429-strategy-for-uk-life-sciences.pdf

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K, Zhao, J. & Mons B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3:160018 DOI: 10.1038/sdata.2016.18.

World Health Organization. (2018). About social determinants of health. Retrieved from http://www.who.int/social_determinants/sdh_definition/en/.