POLITECNICO SETUBAL

Comissão de Ética

PARECER Nº 73/ 2023

Sobre o uso de dados públicos e secundários em estudos e submissão à Comissão de Ética

Pedido de parecer:

"Enquanto coordenadora do curso de Mestrado (…) e sabendo que alguns estudantes vão desenvolver os seus estudos a partir de amostras disponíveis em bancos de dados públicos ou já recolhidos ao longo de vários anos nos serviços onde exercem a sua atividade - estudos esses que entendo como secundários e que dispensam aprovação pela CE-IPS -, pergunto se haverá na mesma algum procedimento a seguir no sentido de a Comissão emitir uma declaração de dispensa para esses estudos."

Fundamentação

Importa clarificar o que se entende por "amostras disponíveis em bancos de dados públicos" e "recolhidos ao longo de vários anos nos serviços onde exercem a sua atividade". Presumivelmente, não se trata de dados com a mesma natureza.

1. Sobre "amostras disponíveis em bancos de dados públicos"

De acordo com a Open Definition, dados abertos "são dados que qualquer pessoa pode aceder, utilizar, modificar e partilhar, para qualquer propósito (sujeito, no máximo, a requisitos que preservem a procedência e a abertura)"[1].

Conforme a Plataforma aberta para dados públicos portugueses[2], representam um subconjunto muito importante do vasto domínio de informação do setor público, cuja reutilização

[1] "Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)." Cf. http://opendefinition.org/

[2] Cf. Plataforma aberta para dados públicos portugueses. https://dados.gov.pt/pt/ "A enorme quantidade de dados que é gerada e centralizada pela Administração Pública congrega em si um enorme potencial de utilização e de desenvolvimentos que podem ser úteis e importantes tanto para o Estado como para a sociedade civil e mundo empresarial. A grande maioria desses dados já são, por lei, considerados públicos. O grande desafio (e a maior preocupação das iniciativas de dados abertos como o dados.gov) passa por facilitar o seu acesso e reutilização, beneficiando vários grupos e sectores da sociedade: • os cidadãos, que passam a ter um acesso mais imediato a informação que lhes pertence por direito, reforçando a visão de transparência e prestação de contas do Estado perante os eleitores; •as instituições governamentais, que se tornam mais transparentes e têm a oportunidade de se tornarem mais eficientes e eficazes, reforçando também o seu papel de serviço público e o próprio acesso a dados de outros organismos;• o setor empresarial, que pode reutilizar informação pública para criar aplicações, plataformas ou serviços com elevado potencial comercial; • e muitos outros setores como o jornalismo, a investigação universitária ou mesmo organizações não-governamentais com preocupações cívicas."

é promovida por uma diretiva europeia[3] (transposta em Portugal pela Lei n.º 26/2016, de 22 de Agosto).

Os dados abertos assentam em princípios da transparência, participação e colaboração. Assim sendo, conjuntos de dados, no contexto dos dados abertos públicos, são agrupamentos de dados em formato digital, dedicados a um tema específico. Por isso distinguem-se *dados abertos* e dados que sejam *apenas* disponibilizados ao público.

Para que os dados possam ser considerados abertos é crucial que não existam restrições ao seu acesso, sejam estas legais ou políticas, tecnológicas ou financeiras – e são sempre abrangidos por licenças abertas, que permitem reutilização comercial; se assim não for, não poderão ser considerados dados abertos.

Dados.gov é o portal nacional de dados abertos da Administração Pública Portuguesa[4], desenvolvido e gerido pela Agência para a Modernização Administrativa, que prossegue as atribuições da Presidência do Conselho de Ministros nas áreas da modernização e simplificação administrativa e da administração eletrónica. A utilização do dados.gov é livre e gratuita.

Outro caso conhecido é o do Instituto Nacional de Estatística, em cujo portal se podem consultar ficheiros de Uso Público. "Estes ficheiros (dados e metainformação) contêm registos anonimizados, tratados e preparados de forma a que a unidade de observação não possa ser identificada direta ou indiretamente, exceto quando se trate de dados estatísticos individuais sobre a Administração Pública. São de acesso gratuito e estão conforme o princípio do segredo estatístico e de proteção de dados pessoais. Este acesso implica a aceitação prévia das condições de utilização."[5]

Constatamos a existência de muitas *bases de dados* disponíveis ao público, como a World Bank Open Data [6], The Global Health Observatory[7], o Google Public Data Explorer[8], entre outros.

Anotamos a diferença na utilização do termo "*bases de dados*" enquanto recursos disponíveis para pesquisa de publicações científicas, plataformas de fornecedores de conteúdos[9];

---

[3] Cf. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en

[4] Portal de dados abertos da Administração Pública https://dados.gov.pt/pt/docs/terms/ . O dados.gov permite: (a) A publicação de dados de interesse público por organismos públicos e todos os outros participantes; (b) A consulta ou o descarregamento de dados por qualquer Utilizador; (c) Discussões sobre conjuntos específicos de dados; (d) Partilha de conjuntos de dados enriquecidos ou alterados; (e) Divulgação de reutilizações de dados abertos. A utilização da plataforma dados.gov está subordinada à aceitação dos termos e condições de utilização. A consulta e descarregamento dos dados disponibilizados em dados.gov não necessita de nenhum tipo de inscrição ou registo prévio. Um participante precisa de se registar e autenticar caso pretenda contribuir para o dados.gov, publicando conjuntos de dados, reutilização, conteúdos, recursos e comentários relativos a conjuntos de dados.

[5] Cf. Portal do INE. Reconhecendo que a comunidade académica "apresenta necessidades especiais no tocante à informação estatística, nomeadamente para o desenvolvimento de trabalhos de investigação e para a elaboração de dissertações de Mestrado e teses de Doutoramento", o Instituto Nacional de Estatística estabeleceu um Protocolo com o Ministério da Educação e Ciência, com o objetivo de facilitar o acesso de investigadoras/es à informação estatística de que necessitam para o exercício da sua atividade, procedendo à credenciação prévia das/os interessadas/os.

[6] Free and open access to global development data. https://data.worldbank.org/

[7] Da OMS - https://www.who.int/data/gho/

[8] Cf. Explorador de Dados Públicos do Google; https://www.google.com/publicdata/directory

[9] Como é o caso da B-On, https://www.b-on.pt/ Biblioteca do conhecimento online. Ou da PubMed (desenvolvido e mantido pelo National Center for Biotechnology Information dos EUA). Ou Web of Science é uma plataforma multidisciplinar, de pesquisa em bases de dados. Ou a ou a WHO Global Index Medicus, https://www.globalindexmedicus.net/

também os *repositórios* podem ter materiais de diversas naturezas, como publicações[10], de bases de dados públicos[11] ou de arquivos[12].

Uma maneira de atenuar (e, até, dirimir) as preocupações éticas do uso de dados pessoais é anonimizá-los por forma a que não se relacionem com pessoas identificáveis. Aliás, é de boa prática que sempre que os fins do tratamento de dados em investigação científica possam ser atingidos com conjuntos de dados que não permitam, ou já não permitam, a identificação dos titulares de dados, devem ser atingidos desse modo. Por anonimização entende-se a utilização de técnicas de conversão de dados pessoais em dados anónimos.

Os dados que não se relacionam com pessoas identificáveis, como dados agregados e estatísticos, não são, em princípio, dados pessoais e estão fora do âmbito do RGPD. Porém, tal só é válido quando o investigador tem acesso apenas aos dados anonimizados, e/ou se o processo de recolha garantir desde logo a anonimização. O que estas bases de dados abertos e/ou publicamente disponíveis têm em comum é o facto de disponibilizarem dados livremente utilizáveis, reutilizáveis e redistribuíveis sem restrições, anonimizados.

"Se o investigador recolher dados pessoais e posteriormente criar um conjunto de dados anonimizados a partir dos primeiros, os novos dados poderão ser considerados ainda dados pessoais, na medida em que o investigador tiver acesso aos dados brutos iniciais. Assim, por exemplo, a criação de um conjunto de dados fruto de informação recolhida junto de participantes através de entrevistas, ainda que posteriormente subtraída de informações de identificação pessoal, pode não traduzir-se em anonimização, até que os dados brutos sejam destruídos ou também anonimizados"[13].

Existem potenciais questões éticas, qualquer que seja a natureza dos dados publicamente disponíveis- podem existir questões de privacidade, de viés, de correção (ou rigor) dos dados [14]. Ainda que esta afirmação pareça afastar-se do propósito original deste parecer, pretendemos chamar a atenção para estes aspetos que podem, eventualmente, passar despercebidos.

Desenvolve-se, atualmente, a chamada "ética dos dados" – tal como "data science", "data ethics". A confiança e a transparência nos processos são tópicos fundamentais na ética dos dados, ainda que pareça existir uma falta de consciência pública dos benefícios, oportunidades, riscos e desafios associados à ciência de dados. São centrais as questões do consentimento e da garantia da privacidade da pessoa e é natural que estes aspetos estejam entrelaçados - análises com foco na privacidade de dados também abordarão questões relativas ao consentimento e responsabilidades profissionais.

---

[10] Como o RCAAP - Repositórios Científicos de Acesso Aberto de Portugal

[11] É o caso do GitHub Cf. https://github.com/devpt-org/public-data-portugal

[12] Como o Arquivo.PT Acesso a conteúdo histórico da Web - https://arquivo.pt/

[13] Orientações aos investigadores sobre proteção de dados pessoais em atividades de investigação científica no ISCTE – Instituto Universitário de Lisboa. P. 6. https://www.iscte-iul.pt/assets/files/2022/06/22/1655919681347_Orientacoes_aos_investigadores_sobre_protecao_de_dados_pessoais.pdf

[14] Sobre o assunto, ver: [1] Cooper AK, Coetzee S. (2020) On the Ethics of Using Publicly-Available Data. *Responsible Design, Implementation and Use of Information and Communication Technology*. 10; 12067:159–71. doi: 10.1007/978-3-030-45002-1_14; [2] Gliniecka, M. (2023). The Ethics of Publicly Available Data Research: A Situated Ethics Framework for Reddit. Social Media + Society, 9(3). https://doi.org/10.1177/20563051231192021; [3] Hand, David J. (2018). Aspects of Data Ethics in a Changing World: Where Are We Now?. *Big Data*.176-190. https://www.liebertpub.com/doi/10.1089/big.2018.0083

Como refere a tomada de posição do Conselho Nacional de Ética para as Ciências da Vida, "no entendimento que a privacidade é um bem cuja proteção deve ser assegurada, defende-se a autodeterminação da informação individual, que atribui ao titular dos dados o seu controlo, decidindo justificadamente os que não podem estar disponíveis, para quem e em circunstância".[15]

Foi proposto um conjunto de princípios básicos de forma a otimizar a reutilização de dados de investigação, a que foi dado o nome de Princípios para Dados FAIR[16]. "Representam um conjunto de normas e boas práticas desenvolvidas pela comunidade para assegurar que os dados ou qualquer objeto digital são Findable (localizáveis), Accessible (acessíveis), Interoperable (interoperáveis) e Re-usable (reutilizáveis)"[17].

2. Sobre amostras disponíveis de dados "já recolhidos ao longo de vários anos nos serviços"

No segundo tópico, considerando a expressão colocada na questão submetida à Comissão de Ética – "dados públicos ou já recolhidos ao longo de vários anos nos serviços onde exercem a sua atividade" – importa ter a certeza que não estamos perante uma situação de uso secundário indevido de dados.

Considera-se uso secundário de dados a sua utilização fora (ou além ou diferente) da finalidade para que foi solicitado e autorizado. "O sistema vale-se da informação obtida para uma finalidade e reutiliza para uma finalidade distinta – por outras palavras, o dado passa de uso primário para uso secundário. Isso torna-o muito mais valioso ao longo do tempo"[18].

Assim, o uso secundário decorre de utilização não relacionada com as finalidades com que o dado foi colhido e tinha sido autorizado – por isso, não existe consentimento para o uso secundário.

Daniel Solove considerou que "o uso secundário é o uso de informações recolhidas para uma finalidade para outra finalidade sem o consentimento do titular dos dados"[19] e colocou-o como atividade danosa no processamento da informação.

---

[15] CNECV (2019). Acesso aos dados de saúde. Tomada de Posição, p. 2. https://www.cnecv.pt/pt/deliberacoes/tomadas-de-posicao/acesso-a-dados-de-saude

[16] Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I.J.; et al., The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* [Internet] 3 (2016), 160018. http://www.nature.com/articles/sdata201618

[17] Manual de Formação em Ciência Aberta. https://foster.gitbook.io/manual-de-formacao-em-ciencia-aberta/02introducaoacienciaaberta/02dados_e_materiais_de_investigacao_abertos . Existem várias formas de tornar os dados de investigação acessíveis, incluindo a publicação dos dados como material suplementar associado a um artigo de investigação, inserir os dados num site disponível publicamente, com os ficheiros disponíveis para download ou depositar os dados num repositório que tenha sido desenvolvido para apoiar a publicação de dados (como o Dataverse, Dryad, Figshare, Zenodo).

[18] Mayer-Schönberger, Viktor; Cukier, Kenneth (2014) Big Data: a revolution that will transform how we live, work, and think. New York, First Mariner Books, p. 103. No original: "The system takes information generated for one purpose and reuses it for another – in other words, the data moves from primary to secondary uses. This makes it much more valuable over time".

[19] Solove, Daniel J. (2006). A Taxonomy of Privacy, 154 U. Pa. L. Rev. 477. https://scholarship.law.upenn.edu/penn_law_review/vol154/iss3/1 "Secondary use is the use of information collected for one purpose for a different purpose without the data subject's consent." (p. 490). A taxonomia que propôs agrupou-se em etapas: a) colheita de informações – i) vigilância; ii) interrogatório; b) processamento de informações – i) agregação; ii) identificação; iii) insegurança; iv) uso secundário; v) exclusão; c) disseminação de informações – i) violação de confidencialidade; ii) divulgação; iii) exposição; iv) aumento na acessibilidade; v) chantagem; vi) apropriação; vii) distorção; d) invasão – i) intrusão; ii) interferência decisional.

"Os usos secundários frustram as expectativas das pessoas sobre como os dados que fornecem serão usados. (…). O uso secundário assemelha-se à quebra de sigilo, na medida em que há uma traição às expectativas da pessoa ao fornecer informações"[20].

O consentimento tem de ser específico, concreto, preciso, dado para um propósito explícito - inclusive, o Regulamento Geral de Proteção de Dados[21] sustenta o princípio da limitação de propósitos, compreendido pela finalidade específica e a compatibilidade, ou seja, o processamento dos dados de forma compatível com a finalidade pela qual os dados foram colhidos.

Assim, dados que foram colhidos, por exemplo, para a prestação de cuidados de saúde, no âmbito clínico, não podem ser utilizados para investigação, a menos que esse uso também tenha sido devidamente autorizado.

Salvaguarde-se que podem estes dados ser irreversivelmente anonimados antes de serem entregues aos investigadores, o que sanaria as referidas questões éticas.

Os dados secundários referem-se a dados colhidos por outra pessoa e pré-existentes em relação ao projeto de pesquisa em causa, também podendo decorrer de investigações anteriores, incluindo publicações de estudos primários[22].

Para pesquisas que dependem exclusivamente do uso secundário de informações anónimas ou anonimizadas, a revisão ética não é necessária[23].

A Comissão de Ética pronuncia-se sobre projetos de investigação clínica[24], previamente à sua implementação, competindo-lhe emitir parecer que considere a adequação científica e ética dos investigadores, a avaliação independente dos aspetos metodológicos, éticos e legais dos estudos de investigação clínica bem como a monitorização e acompanhamento dos estudos de investigação clínica que decorrem na instituição[25].

---

[20] Idem, p. 490 "Secondary uses thwart people's expectations about how the data they give out will be used. (…) Secondary use resembles breach of confidentiality, in that there is a betrayal of the person's expectations when giving out information."

[21] RGPD – 32. O consentimento do titular dos dados deverá ser dado mediante um ato positivo claro que indique uma manifestação de vontade livre, específica, informada e inequívoca de que o titular de dados consente no tratamento dos dados que lhe digam respeito, como por exemplo mediante uma declaração escrita, inclusive em formato eletrónico, ou uma declaração oral. O consentimento pode ser dado validando uma opção ao visitar um sítio web na Internet, selecionando os parâmetros técnicos para os serviços da sociedade da informação ou mediante outra declaração ou conduta que indique claramente nesse contexto que aceita o tratamento proposto dos seus dados pessoais. O silêncio, as opções pré-validadas ou a omissão não deverão, por conseguinte, constituir um consentimento. O consentimento deverá abranger todas as atividades de tratamento realizadas com a mesma finalidade. Nos casos em que o tratamento sirva fins múltiplos, deverá ser dado um consentimento para todos esses fins. Se o consentimento tiver de ser dado no seguimento de um pedido apresentado por via eletrónica, esse pedido tem de ser claro e conciso e não pode perturbar desnecessariamente a utilização do serviço para o qual é fornecido."

[22] Salientamos que dados primários são os dados obtidos diretamente do processo de investigação, instrumento ou metodologia científica, sem que tenham sofrido qualquer processamento ou transformação. Dados secundários são resultantes da interpretação, processamento ou transformação de dados primários.

[23] Ainda assim, sugere-se aos investigadores que considerem algumas questões como: (a) Onde foram obtidos os dados? De um protocolo previamente aprovado? Repositório ou arquivo de dados regulamentados? (b) Há alguma restrição ao uso ou divulgação dos resultados? (c) Algum potencial para ligação de dados?

[24] "considera-se investigação clínica a investigação conduzida em seres humanos ou em material de origem humana, tais como tecidos, espécimes e fenómenos cognitivos, para os quais um investigador interage diretamente com seres humanos." Decreto lei nº 80/2018 de 15 de outubro, artigo 1º, nº 2.

[25] Decreto lei nº 80/2018 de 15 de outubro. Estabelece os princípios e regras aplicáveis às comissões de ética que funcionam nas instituições de saúde, nas instituições de ensino superior e em centros de investigação biomédica que desenvolvam investigação clínica. Ver artigo https://files.dre.pt/1s/2018/10/19800/0496504970.pdf

Conclusão

É nosso entendimento que:

1. A revisão ética é parte fundamental do processo de investigação e promove a proteção dos participantes, dos investigadores assim como a integridade da produção científica. São elegíveis para parecer ético qualquer tipo de estudo primário - ou seja, estudo empírico que envolva recolha de dados com participantes, em que os dados são colhidos ou as variáveis observadas e analisadas, com ou sem implementação de intervenções.
2. Os dados disponíveis em bases de dados abertos e/ou publicamente disponíveis têm em comum o facto de serem localizáveis, acessíveis, interoperáveis, livremente utilizáveis, reutilizáveis e redistribuíveis sem restrições. Por via de regra, estão anonimizados. Não obstante, recomenda-se que os investigadores procurem identificar as fontes de informação assim como os termos de utilização.
3. Para dados secundários fornecidos por terceiros, importa garantir que se encontram irreversivelmente anonimizados, assegurada a conversão de dados pessoais em dados anónimos.
4. Para os estudos que não envolvam participantes, estudos secundários que utilizam dados já disponíveis em bases de dados públicas para selecionar as evidências, não se requer aprovação pela CE-IPS nem se justifica declarar a sua dispensa.

Aprovado por unanimidade em reunião plenária
18 dezembro 2023.

Presidente da Comissão de Ética do IPS

Assinado por: **Lucília Rosa Mateus Nunes**
Num. de Identificação: 06064421