

# Big Data em Pesquisas Populacionais sobre Câncer: Desafios e Oportunidades

doi: <https://doi.org/10.32635/2176-9745.RBC.2019v65n4.718>

*Big Data in Population Cancer Research: Challenges and Opportunities*

*Big Data en la Investigación de Cáncer de Población: Desafíos y Oportunidades*

Luiz Claudio Santos Thuler<sup>1</sup>; Zulmira Hartz<sup>2</sup>

## INTRODUÇÃO

Nos últimos anos, a popularização dos computadores, o crescente acesso à Internet, a geração de uma quantidade astronômica de dados, ativa ou passivamente originados, têm contribuído para o surgimento exponencial de grandes bancos de dados, os *Big Data*. Eles suscitam um novo desafio que é o de identificar maneiras de obter e extrair informações de forma eficiente e aproveitar as oportunidades oferecidas pelos avanços tecnológicos<sup>1</sup>. Do ponto de vista etimológico, as informações são originárias de coleções de dados, e constituem por sua vez a base do conhecimento. Entretanto, está cada vez mais difícil separar dados de informações, visto que são, muitas vezes, intercambiáveis<sup>2</sup>.

O termo *Big Data* foi usado pela primeira vez por Michael Cox e David Ellsworth, cientistas da *National Aeronautics and Space Administration* (NASA), em um artigo escrito em 1997, expondo a dificuldade de trabalhar com conjuntos muito grandes de dados<sup>3</sup>. Embora existam inúmeras definições para *Big Data*, o que nos conduz ao raciocínio de que nenhuma delas é *per se* satisfatória, em geral o termo tem sido usado quando estão presentes três Vs: volume, variedade e velocidade. Esses três Vs, propostos inicialmente por Laney<sup>4</sup>, vêm sendo amplamente aceitos como as principais dimensões atinentes aos *Big Data*. Posteriormente, alguns Vs foram adicionados: veracidade, versatilidade e viabilidade; e a combinação de todos os Vs geraria valor<sup>1,5,6</sup>. Entretanto, a busca por Vs não cessou, tendo sido, recentemente, sugeridos por Shafer<sup>7</sup> 42 Vs.

Berman (apud Jackson<sup>8</sup>) propõe que o que diferencia os *Big Data* dos *Small Data* é que os últimos geralmente estão contidos em um único computador, estão localizados dentro de uma organização, são geralmente bem estruturados e formatados (por exemplo, em planilhas) e

sua coleta normalmente atende a um propósito específico. Já os *Big Data*, não raro, estão distribuídos por vários servidores de Internet, podem ser estruturados ou não (por exemplo, texto livre, imagens, vídeos, áudios), provenientes de diversas fontes e gerados por muitas pessoas, não tendo sido projetados para o propósito específico de determinada pesquisa<sup>8</sup>.

No que diz respeito à geração de dados em saúde, Herland et al.<sup>9</sup> sugerem que o uso dos *Big Data* deve considerar dois níveis: a questão de pesquisa e a coleta dos dados. As questões epidemiológicas, por exemplo, devem apoiar-se em dados populacionais; já as questões clínicas podem apoiar-se em dados moleculares (como, por exemplo, o uso de dados de expressão gênica para fazer previsões clínicas, ou os subtipos de câncer obtidos por exames anatomopatológicos, ou ainda o uso de mensagens para ajudar os pacientes a obterem informações médicas), ou dados coletados diretamente de pacientes (por exemplo, modelos preditivos visando a identificar fatores prognósticos); por sua vez, questões relacionadas à biologia humana podem se utilizar de dados teciduais (exames laboratoriais ou de imagem). Já em relação à coleta de dados, idealmente, devem ser consideradas todas as possíveis fontes para a sua coleta. Os métodos mais usados nessa fase são arquivos, rastreadores da *Web* e *softwares* de monitoramento de rede. Na sequência, os dados devem ser processados e armazenados. No caso de dados provenientes de diferentes fontes, devem ser primeiramente integrados e, a seguir, examinados quanto à acurácia, completude e consistência. Para sua análise, devem ser definidas as métricas, algoritmos e as formas de visualização. Um dos maiores desafios tem sido identificar ferramentas que suportem seu armazenamento e gerenciamento (como, por exemplo: *Hadoop Distributed File System* – HDFS), volume (*Oracle NoSQL*, *Apache HBase*, *Apache Cassandra*, *Apache Hive*), capacidade de processamento (*MapReduce*,

<sup>1</sup> Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA). Universidade Federal do Estado do Rio de Janeiro (Unirio). Rio de Janeiro (RJ), Brasil. Orcid iD: <https://orcid.org/0000-0003-2550-6537>

<sup>2</sup> *Global Health and Tropical Medicine* (GHTM). Instituto de Higiene e Medicina Tropical (IHMT) da Universidade Nova de Lisboa (UNL). Lisboa, Portugal. Orcid iD: <https://orcid.org/0000-0001-9780-9428>

**Endereço para correspondência:** Luiz Claudio Santos Thuler. Divisão de Pesquisa Clínica do INCA. Rua André Cavalcanti, 37 – Centro. Rio de Janeiro (RJ), Brasil. CEP 20231-050. E-mail: [lthuler@gmail.com](mailto:lthuler@gmail.com)



*Apache Hadoop*), capacidade de integração (*Oracle Big Data Connectors, Oracle Data Integrator*) e capacidade de análise estatística (*R and Oracle R Enterprise*). Por fim, com base nos resultados, poderão ser tomadas decisões para reforçar comportamentos positivos ou alterar aqueles indesejados<sup>1</sup>.

## BIG DATA E CÂNCER

Diversos autores têm utilizado os *Big Data* para responder a questões específicas na área de câncer. Meyer et al.<sup>10</sup> propõem o uso de *Big Data* para a pesquisa de base populacional em câncer e citam diversos autores que têm se utilizado do Sistema Integrado para Informação e Vigilância do Câncer da Carolina do Norte (do inglês: *The Integrated Cancer Information and Surveillance System - ICISS*) e do sistema de seguros de saúde *Medicare*. Esses autores destacam, por exemplo, o artigo de Wheeler et al.<sup>11</sup>, no qual são discutidas as desigualdades na utilização da radioterapia após cirurgia conservadora em mulheres de baixa renda com câncer de mama em estágio inicial. Nesse estudo, somente 75% das mulheres iniciaram radioterapia dentro de seis meses após o diagnóstico do câncer de mama e, embora a oferta de radioterapia tenha melhorado ao longo dos anos, seu início dentro de um ano foi mais frequente nas pacientes com menos comorbidades e em uso de quimioterapia. Já o estudo de Holmes et al.<sup>12</sup>, também apresentado por Meyer et al.<sup>10</sup>, mostrou que pacientes que viviam mais longe de um urologista eram mais propensos a serem diagnosticados com câncer de próstata de alto risco. Os autores enfatizam que, desproporcionalmente, a distância impactou mais negativamente homens negros do que brancos: entre aqueles que viviam a mais de 32 quilômetros do urologista mais próximo, 49% dos pacientes brancos e 79% dos pacientes negros foram diagnosticados com doença de alto risco.

Em um artigo sobre o uso de *Big Data* para avaliação da qualidade da atenção em oncologia, Broughman et al.<sup>13</sup> propõem que, para serem considerados de alta qualidade, os serviços devem ser: seguros, efetivos, centrados no paciente, oportunos, eficientes e equitativos. Os autores citam 26 estudos que utilizam os dados do Programa de Vigilância, Epidemiologia e Resultados Finais (do inglês: *The Surveillance, Epidemiology, and End Results - SEER*) do Instituto Nacional de Câncer dos Estados Unidos e da Base de Dados Nacional sobre Câncer (do inglês: *National Cancer Database - NCDDB*) com essa finalidade. Nesses estudos, o número de pacientes variou entre 1.962 e 2.342.558, sendo destacado, por exemplo, o estudo de Fedewa et al.<sup>14</sup>, que analisou 312.339 homens quanto à gravidade do câncer de próstata ao diagnóstico. Pacientes afro-americanos, hispânicos e asiáticos, em comparação

com os caucasianos, bem como aqueles sem seguro e beneficiários dos programas federal e estadual de ajuda com os custos médicos disponibilizados para pessoas com renda e recursos limitados (*Medicaid*), em comparação com pacientes com seguro privado, eram mais propensos a apresentar doença mais agressiva ao diagnóstico. De forma semelhante, citam o estudo elaborado por Gross et al.<sup>15</sup>, no qual foram usados dados do SEER e do Medicare para mostrar que pacientes afro-americanos *versus* caucasianos tinham menos probabilidade de receber tratamento para câncer de pulmão (cirurgia), mama (radioterapia), cólon (terapia adjuvante) e próstata (prostatectomia, braquiterapia ou radioterapia).

A maior iniciativa público-privada da Europa, a Iniciativa sobre Medicamentos Inovadores 2 (do inglês: *The Innovative Medicines Initiative 2 - IMI2*), empreendimento conjunto entre a União Europeia e a Federação Europeia de Associações e Indústrias Farmacêuticas (do inglês: *European Federation of Pharmaceutical Industries and Associations - EFPIA*), lançou recentemente o programa *Big Data* para Melhores Resultados (do inglês: *Big Data for Better Outcomes - BD4BO*). O programa visa a catalisar e a apoiar a evolução de sistemas de cuidados de saúde sustentáveis orientados por resultados. Na área de câncer, o objetivo principal é aumentar o corpo de evidências visando a melhorar os resultados. Por meio de medidas epidemiológicas, clínicas, econômicas relatadas pelos pacientes (do inglês: *Patient Reported Outcomes - PRO*), serão analisados os fatores preditivos de complicações e o risco de desenvolvimento de efeitos adversos em todos os estágios da doença<sup>16</sup>.

## BIG DATA E CÂNCER NO BRASIL

No Brasil, as informações sobre câncer têm sido obtidas de distintas fontes, conforme sua especificidade.

O Sistema de Informações sobre Mortalidade (SIM) tem dados disponíveis dos óbitos por todas as causas, incluindo o câncer, desde 1979, e vem sendo amplamente utilizado, particularmente para analisar as tendências de mortalidade pelos diferentes tipos de câncer no país. A partir desses dados, indicadores e taxas de mortalidade por câncer no Brasil são disponibilizados pelo Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA)<sup>17</sup> em forma de tabelas, gráficos e mapas no *Atlas On-line de Mortalidade*.

Dados de incidência do câncer podem ser obtidos a partir dos Registros de Câncer de Base Populacional (RCBP), que têm por objeto coletar dados de todos os casos incidentes de câncer residentes em uma área geográfica delimitada, ocorridos em um determinado período de tempo. Os primeiros RCBP foram implantados em Recife (1967) e em São Paulo (1969)<sup>18</sup>. Atualmente,

existem mais de 20 RCBP em atividade, com pelo menos cinco anos de informações consolidadas, cobrindo cerca de ¼ da população do país<sup>19</sup>. Seus dados têm sido utilizados na análise da incidência de diversos tipos de câncer e em populações específicas.

Por sua vez, os Registros Hospitalares de Câncer (RHC) concentram dados referentes aos pacientes matriculados em um único hospital, sendo úteis para avaliação da qualidade da assistência prestada. No Brasil, o primeiro RHC foi criado em 1982, no INCA, iniciando a coleta de dados a partir de janeiro de 1983<sup>19</sup>. Atualmente, um tabulador de dados (Integrador RHC) disponibiliza na Internet um acervo de mais de três milhões de casos de câncer, cobrindo as 27 Unidades da Federação. Os dados concentram-se entre os anos 2000 e 2017 (99,9% dos casos) e são provenientes de mais de 330 hospitais, com um número de casos por hospital variando de 17 a 158.516 (dados atualizados em 29 de outubro de 2019). Diversos estudos têm utilizado os dados do Integrador RHC para responder a questões envolvendo desde os fatores de risco para o câncer até seu tratamento e desfechos.

Estão ainda disponíveis informações sobre exames citopatológicos, anatomopatológicos e mamografias realizados pelo Sistema Único de Saúde (SUS), disponibilizados pelo Sistema de Informação do Câncer do Colo do Útero (Siscolo) e pelo Sistema de Informação do Câncer de Mama (Sismama), respectivamente. Os dados do Siscolo e Sismama, reunidos recentemente no Sistema de Informação do Câncer (Siscan), têm sido utilizados para o monitoramento das diretrizes, objetivos, metas e indicadores municipais, regionais, estaduais e nacionais do SUS, pactuados na comissão intergestores tripartite<sup>20</sup>. Os dois indicadores específicos para câncer são a razão entre o número de exames citopatológicos do colo do útero em mulheres de 25 a 64 anos e a população da mesma faixa etária, e a razão entre o número de exames de mamografia de rastreamento realizados em mulheres de 50 a 69 anos e a população da mesma faixa etária. A edição 2015 (última disponível em outubro de 2019) apresenta a série histórica de 2008 a 2015, permitindo acompanhar a tendência de cada indicador no território analisado.

Devem ser destacados ainda os bancos de dados não específicos para câncer, mantidos pelo Ministério da Saúde, como por exemplo o Sistema de Informações sobre Nascidos Vivos (Sinasc), o Sistema de Informação de Agravos de Notificação (Sinam), o Sistema de Informações de Atenção Básica (Siab), o Sistema de Informações Ambulatoriais (SIA), o Sistema de Informações Hospitalares (SIH), a Autorização de Procedimentos Ambulatoriais de Alta Complexidade/Custo (Apac), bem como os bancos de dados de inquéritos e pesquisas disponibilizados no sítio do Departamento

de Informática do Sistema Único de Saúde do Brasil (DATASUS), e que podem ser úteis conforme o objeto do estudo<sup>21</sup>.

Embora essas bases de dados não cumpram necessariamente todas as dimensões desejáveis aos *Big Data*, como velocidade (alguns dados são disponibilizados entre um e três anos após sua ocorrência), outras características como volume, variedade, veracidade, versatilidade, viabilidade, em geral, estão presentes.

## EXEMPLOS DO USO DE *BIG DATA* EM CÂNCER NO BRASIL

Em um estudo pioneiro com o Integrador RHC, Thuler et al.<sup>22</sup>, após analisarem 77.317 casos de câncer do colo do útero, apontaram para o elevado percentual de casos cujo diagnóstico ocorreu quando a doença se encontrava em estágio avançado. Mais de 1/3 dos casos apresentavam-se nos estádios III e IV no momento do diagnóstico. Os autores destacaram ainda que o percentual de casos com doença avançada aumentou de 32,8% para 35,5% entre o primeiro (2000-2004) e o segundo (2005-2009) quinquênios analisados. Em outro estudo, também com dados do Integrador RHC, dessa feita envolvendo 59.317 mulheres com câncer de mama, Abrahão et al.<sup>23</sup> identificaram fatores associados ao diagnóstico em estágio  $\geq$ IIB. A análise ajustada mostrou que ter idade  $<$ 50 anos, apresentar baixa escolaridade e residir nas regiões geográficas mais carentes do país estavam associados à doença avançada no momento do diagnóstico.

A incidência, a morbidade e a mortalidade do câncer de pulmão entre 2000 e 2014 foram analisadas a partir de três bases de dados: RCBP, Integrador RHC e SIM, respectivamente. Costa et al.<sup>24</sup> mostram que, enquanto as taxas de incidência e mortalidade por câncer de pulmão diminuíram nos homens, em mulheres essas aumentaram. Chamaram ainda atenção para a redução da razão homem:mulher ao longo do período. Usando as mesmas bases de dados, Melo et al.<sup>25</sup> realizaram o maior estudo conduzido no Brasil sobre melanoma cutâneo. Verificaram que o aumento nas taxas de incidência observado em ambos os sexos no período se acompanhou de taxas de mortalidade praticamente estáveis apontando para possíveis melhorias no diagnóstico e tratamento dos casos. Chamaram a atenção ainda para o fato de que quase metade dos pacientes teve o diagnóstico em estádios avançados da doença.

Os grandes bancos de dados nacionais também têm sido usados para responder a perguntas referentes aos cânceres infantjuvenis. Em um estudo visando a analisar as tendências do câncer do colo do útero em adolescentes e mulheres jovens (15 a 29 anos), Viana et al.<sup>26</sup> utilizaram dados de 21 RCBP e concluíram que, no Brasil, as taxas de incidência são elevadas e encontram-se em ascensão

ou estáveis na maior parte dos RCBP, apontando para a necessidade de intervenções mandatórias dirigidas à prevenção e ao controle da doença nesse grupo etário específico. Por sua vez, Reis et al.<sup>27</sup> utilizaram dados de 12 RCBP para analisar o risco de desenvolver leucemias na primeira infância (crianças ≤5 anos de idade). Ao propor um estudo caso-coorte, 272 pacientes com leucemia foram comparados a 1.088 controles identificados, a partir do banco de dados do Sinasc, na mesma faixa de idade, localidade e período (2000-2009). Observou-se, nas crianças com malformação congênita e naquelas cujas mães trabalhavam expostas a produtos químicos (agropecuária, química e indústria petroquímica), maior risco de desenvolver leucemia.

O padrão de anormalidades cervicais no Estado do Maranhão foi analisado por Pinho-França et al.<sup>28</sup> utilizando os dados do Siscolo. A avaliação dos resultados da citologia cervical de 1.363.689 mulheres examinadas entre 2007 e 2012 mostrou que quanto pior o Índice de Desenvolvimento Humano (IDH) do município onde a mulher reside, maior o risco de apresentar lesão intraepitelial escamosa de alto grau ou câncer do colo do útero. Os autores destacaram que, para o controle da doença, faz-se necessário reduzir as desigualdades sociais e melhorar a disponibilidade de serviços de saúde. Por outro lado, analisando a associação entre o IDH e o estágio do câncer do colo do útero ao diagnóstico, Vale et al.<sup>29</sup> mostraram, a partir de 51.158 casos registrados no Integrador RHC, haver correlação inversa entre o IDH da cidade onde a mulher reside e o estágio da doença. Os autores chamaram a atenção para o elevado percentual de casos diagnosticados em estádios avançados e apontaram a necessidade de estratégias regionais que melhorem a eficiência das intervenções dirigidas ao câncer em países com grandes disparidades socioeconômicas, como é o caso do Brasil.

Com a entrada em vigor da Lei 13.685<sup>30</sup>, em 24 de dezembro de 2018, por meio da qual doenças, agravos e eventos em saúde relacionados às neoplasias atendidos nos serviços de saúde públicos e privados do país terão que ser notificados compulsoriamente, ampliam-se as perspectivas no que diz respeito à completude dos dados de câncer no país. Cumpre destacar que essa Lei carece de regulamentação.

## DESAFIOS E OPORTUNIDADES NO USO DE *BIG DATA* EM PESQUISAS POPULACIONAIS SOBRE CÂNCER

Embora se vislumbre um potencial imenso na utilização dos *Big Data* na área da saúde, há peculiaridades que devem ser consideradas. A interpretação de resultados de testes estatísticos com grandes volumes de dados deve ser cautelosa, já que ficam aumentadas as probabilidades de se identificar associações como estatisticamente significantes,

mas que, muitas vezes, não podem ser definidas como associações causais ou clinicamente importantes<sup>31,32</sup>. Por outro lado, a quantidade de variáveis existentes nos bancos de dados amplia as chances de se identificar associações espúrias obtidas a partir de análises desvinculadas de modelos lógicos e/ou teóricos. Como estratégias, tem-se recomendado evitar a realização de testes de hipótese para todas as variáveis disponíveis, limitando-se apenas aquelas para as quais existe fundamentação teórica, empregando estratégias para reduzir a ocorrência de erro tipo I (quando o resultado do teste apresenta significância estatística, mas na verdade sua ocorrência se deu por acaso) como o teste de Bonferroni, a correção de Bonferroni e o controle das taxas de falsas descobertas por meio do procedimento de Benjamini-Hochberg<sup>21,33</sup>. Outras questões que não podem ser deixadas de lado referem-se à validade de construto e à confiabilidade<sup>34</sup>. Nesse sentido, além das dimensões necessárias aos *Big Data* mais frequentemente descritas como velocidade, volume, variedade, veracidade, versatilidade, viabilidade, torna-se essencial o V de validade, proposto por Normandeau<sup>35</sup>. Cumpre ainda destacar que, uma vez utilizados com a finalidade de pesquisa, os *Big Data* não estão isentos do cumprimento dos requisitos éticos para pesquisas envolvendo seres humanos, sendo necessário garantir a privacidade, a confidencialidade, a segurança e a integridade dos dados<sup>21,36</sup>.

## CONCLUSÃO

Nesse contexto, consideradas as potencialidades no uso dos *Big Data* em pesquisas populacionais sobre câncer, é fundamental dar utilidade e visibilidade aos dados sobre câncer disponibilizados por órgãos públicos.

## CONTRIBUIÇÕES

Ambos os autores contribuíram igualmente de forma substancial em todas as etapas do trabalho e aprovaram a versão final a ser publicada.

## DECLARAÇÃO DE CONFLITO DE INTERESSES

Nada a declarar.

## FONTES DE FINANCIAMENTO

Não há.

## REFERÊNCIAS

1. Singh MK, Kumar DG, editors. Effective big data management and opportunities for implementation [Internet]. Pennsylvania, USA: IGI Global; 2016.

- Chapter 1, Big data: challenges, opportunities and realities; [cited 2019 Sept 17]; p. 1-24. Available from: <https://arxiv.org/ftp/arxiv/papers/1705/1705.04928.pdf>
- The Economist [Internet]. London: The Economist Newspaper; [1997]. Data, data everywhere; [2010 Feb 27]. [cited 2019 Sept 17]. Available from: <https://www.economist.com/special-report/2010/02/27/data-data-everywhere>.
  - Cox M, Ellsworth D. Managing big data for scientific visualization. ACM Siggraph [Internet]. 1997 [cited 2019 Sept 17];97:21-38. Available from: [https://www.researchgate.net/publication/238704525\\_Managing\\_big\\_data\\_for\\_scientific\\_visualization](https://www.researchgate.net/publication/238704525_Managing_big_data_for_scientific_visualization)
  - Laney D. 3D data management: controlling data volume, velocity, and variety [Internet]. Stamford: META Group; 2001 Feb 6. [cited 2019 Sept 17]. File: 949. Available from: <https://blogs.gartner.com/douglaney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>
  - Dias JA, Duarte P. Big data opportunities in healthcare. How can medical affairs contribute? Rev Port Farmacoter. 2015;7(4):230-236. doi: <https://doi.org/10.25756/rpf.v7i4.99>
  - Magalhães J, Hartz Z, Martins MRO. Big Data para a investigação em saúde e a ciência aberta: um contributo para a gestão do conhecimento. An Inst Hig Med Trop. 2016;15(Supl. 2):S75-S82. doi: <https://doi.org/10.25761/anaisihmt.130>
  - KDnuggets [Internet]. [place unknown]: KDnuggets; 1997. Shafer T. The 42 V's of big data and data science; 2017. [cited 2019 Sept 17]. Available from: <https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html>
  - Jackson S. Big data monitoring and evaluation: a theoretical framework, tools, and lessons learned from practice [Internet]. [Draft v2.0] Jakarta: United Nations Global Pulse; 2015 Dec. [cited 2019 Sept 17]. Available from: [http://unglobalpulse.org/sites/default/files/Annex%201%20Big\\_data\\_monitoring\\_and\\_evaluation.pdf](http://unglobalpulse.org/sites/default/files/Annex%201%20Big_data_monitoring_and_evaluation.pdf)
  - Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. J Big Data. 2014;1(2). doi: <https://doi.org/10.1186/2196-1115-1-2>
  - Meyer AM, Olshan AF, Green L, et al. Big data for population-based cancer research: the integrated cancer information and surveillance system. N C Med J. 2014 Jul-Aug;75(4):265-9. doi: <https://doi.org/10.18043/ncm.75.4.265>
  - Wheeler SB, Wu Y, Meyer AM, et al. Use and timeliness of radiation therapy after breast-conserving surgery in low-income women with early-stage breast cancer. Cancer Invest. 2012;30(4):258-267. doi: <https://doi.org/10.3109/07357907.2012.658937>
  - Holmes JA, Carpenter WR, Wu Y, et al. Impact of distance to a urologist on early diagnosis of prostate cancer among black and white patients. J Urol. 2012;187(3):883-888. doi: <https://doi.org/10.1016/j.juro.2011.10.156>
  - Broughman JR, Chen RC. Using big data for quality assessment in oncology. J Comp Eff Res. 2016;5(3):309-19. doi: <https://doi.org/10.2217/cer-2015-0021>
  - Fedewa SA, Etzioni R, Flanders WD, et al. Association of insurance and race/ethnicity with disease severity among men diagnosed with prostate cancer, National Cancer Database 2004-2006. Cancer Epidemiol Biomarkers Prev. 2010 Oct;19(10):2437-44. doi: <https://doi.org/10.1158/1055-9965.EPI-10-0299>
  - Gross CP, Smith BD, Wolf E, et al. Racial disparities in cancer therapy: did the gap narrow between 1992 and 2002?. Cancer. 2008 Feb 15;112(4):900-8. doi: <https://doi.org/10.1002/cncr.23228>
  - Szócska M, Jayawardana S, Smand C, et al. Big data for better outcomes: supporting health care system transformation in Europe. Eurohealth [Internet]. 2017. [cited 2019 Sept 17];23(1):7-9. Available from: [http://www.euro.who.int/\\_\\_data/assets/pdf\\_file/0004/335812/Eurohealth-V23-N1-2017.pdf?ua=1](http://www.euro.who.int/__data/assets/pdf_file/0004/335812/Eurohealth-V23-N1-2017.pdf?ua=1)
  - Instituto Nacional de Câncer José Alencar Gomes da Silva. Atlas on-line de mortalidade [Internet]. Rio de Janeiro: INCA; [2014] - [atualizado 2019 maio 30; acesso 2019 set. 17]. Disponível em: <https://mortalidade.inca.gov.br/MortalidadeWeb/>
  - Ministério da Saúde (BR), Departamento de Ciência e Tecnologia, Secretaria de Ciência e Tecnologia e Insumos Estratégicos. Integração de informações dos registros de câncer brasileiros. Rev Saúde Publica. 2007 Oct;41(5):865-8. doi: <http://dx.doi.org/10.1590/S0034-89102007000500024>
  - Instituto Nacional de Câncer José Alencar Gomes da Silva. Informação dos registros hospitalares de câncer como estratégia de transformação: perfil do Instituto Nacional de Câncer José Alencar Gomes da Silva em 25 anos [Internet]. Rio de Janeiro: INCA; 2012. [acesso 2019 set. 17]. 100 p. Disponível em: <https://www.inca.gov.br/publicacoes/livros/informacao-dos-registros-hospitalares-de-cancer-como-estrategia-de-transformacao>
  - Ministério da Saúde (BR); Comissão Intergestores Tripartite. Resolução nº 5, de 19 de junho de 2013. Diário Oficial da União, Brasília, DF; 2013 jun. 20. Seção I, p. 63.
  - Chiavegatto Filho ADP. Uso de big data em saúde no Brasil: perspectivas para um futuro próximo. Epidemiol Serv Saúde. 2015;24(2):325-332. doi: <http://dx.doi.org/10.5123/S1679-49742015000200015>
  - Thuler LCS, Bergmann A, Casado L. Perfil das pacientes com câncer do colo do útero no Brasil, 2000-2009: estudo de base secundária. Rev Bras Cancerol. 2012;58:351-357. doi: <https://doi.org/10.32635/2176-9745.RBC.2012v58n3.583>

23. Abrahão KS, Bergmann A, Aguiar SS, et al. Determinants of advanced stage presentation of breast cancer in 87,969 Brazilian women. *Maturitas*. 2015 Dec;82(4):365-70. doi: <https://doi.org/10.1016/j.maturitas.2015.07.021>
24. Costa GJ, Mello MJG, Ferreira CG, et al. Increased incidence, morbidity and mortality rates for lung cancer in women in Brazil between 2000 and 2014: an analysis of three types of sources of secondary data. *Lung Cancer*. 2018 Nov;125:77-85. doi: <https://doi.org/10.1016/j.lungcan.2018.09.005>
25. Melo AC, Wainstein AJA, Buzaid AC, et al. Melanoma signature in Brazil: epidemiology, incidence, mortality, and trend lessons from a continental mixed population country in the past 15 years. *Melanoma Res*. 2018;28(6):629-36. doi: <https://doi.org/10.1097/CMR.0000000000000511>
26. Viana LS, Balmant NV, Silva NP, et al. Incidence Trends of Cervical Cancer in Adolescents and Young Adults: Brazilian Population Based Data. *J Adolesc Young Adult Oncol*. 2018 Feb;7(1):54-60. doi: <https://doi.org/10.1089/jayao.2017.0048>
27. Reis RS, Silva NP, Santos MO, et al. Mother and child characteristics at birth and early age leukemia: a case-cohort population-based study. *J Pediatr*. 2017 Nov-Dec;93(6):610-18. doi: <http://dx.doi.org/10.1016/j.jpeds.2016.12.009>
28. Pinho-França JR, Chein MBC, Thuler LCS. Patterns of cervical cytological abnormalities according to the Human Development Index in the northeast region of Brazil. *BMC Womens Health*. 2016 Aug 12;16:54. doi: <http://dx.doi.org/10.1186/s12905-016-0334-2>
29. Vale DB, Sauvaget C, Muwonge R, et al. Level of human development is associated with cervical cancer stage at diagnosis. *J Obstet Gynaecol*. 2019;39(1):86-90. doi: <http://dx.doi.org/10.1080/01443615.2018.1463976>
30. Presidência da República (BR). Lei nº 13.685, de 25 de junho de 2018. *Diário Oficial da União, Brasília, DF*; 2018 jun. 26. Seção I, p. 1.
31. Groeneveld PW, Rumsfeld JS. Can big data fulfill its promise? *Circ Cardiovasc Qual Outcomes*. 2016;9(6):679-82. doi: <http://dx.doi.org/10.1161/CIRCOUTCOMES.116.003097>
32. Kamper SJ. Interpreting Outcomes 2-Statistical Significance and Clinical Meaningfulness: Linking Evidence to Practice. *J Orthop Sports Phys Ther*. 2019 Jul;49(7):559-60. doi: <http://dx.doi.org/10.2519/jospt.2019.0704>
33. McDonald JH. *Handbook of Biological Statistics* [Internet]. 3rd ed. Baltimore: Sparky House Publishing; 2014 [revised 2015 July 20]. Maryland, multiple comparisons; [cited 2019 Sept 17]; [about I screen]. Available from: <http://www.biostathandbook.com/multiplecomparisons.html>
34. Lazer D, Kennedy R, King G, et al. The parable of google flu: traps in big data analysis. *Science* [Internet]. 2014 [cited 2019 Sept 17];343:1203-1205. Available from: <https://gking.harvard.edu/files/gking/files/0314policyforumff.pdf>
35. insideBIGDATA [Internet]. Portland, OR: insideBIGDATA; [2011]. Normandeau K. Beyond volume, variety and velocity is the issue of big data veracity; 2013 Sept 12. [cited 2019 Sept 17]. Available from: <https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/>
36. Salerno J, Knoppers BM, Lee LM, et al. Ethics, big data and computing in epidemiology and public health. *Ann Epidemiol*. 2017;27(5):297-301. doi: <http://dx.doi.org/10.1016/j.annepidem.2017.05.002>

Recebido em 30/10/2019  
Aprovado em 1/11/2019