Bancos de dados, *big data* e inteligência artificial: o que os profissionais de saúde precisam saber sobre eles*

Leticia Leone Lauricella[I], *Paulo Manuel Pêgo-Fernandes*[II]

Instituto do Coração (InCor), Hospital das Clínicas HCFMUSP, Faculdade de Medicina, Universidade de Sao Paulo, Sao Paulo, SP, BR

"A informação é o petróleo do século 21, e a análise é o motor de combustão", disse Peter Sondergaard, vice-presidente sênior da Gartner Research. Quanto mais informações tivermos, maior a probabilidade de encontrarmos correlações que não são óbvias aos olhos e que podem mudar completamente a maneira como pensamos ou agimos. Estamos vivendo uma realidade na qual toneladas de dados de várias fontes diferentes são coletadas rotineiramente, muitas vezes sem que percebamos. Todos os aspectos de nossas vidas, como atividades sociais, padrões de consumo, pesquisas na internet, deslocamentos geográficos em sistemas de posicionamento global (GPS) e questões de saúde estão, de alguma forma, sendo transformados em dados. Este grande volume de novos dados sendo gerado está sobrecarregando a capacidade das instituições de gerenciá-los e dos pesquisadores de utilizá-los, uma situação que tem sido chamada de "dilúvio de dados".[1]

Uma das áreas mais notáveis em que a análise de dados está causando grandes mudanças é na área da saúde. Além da pesquisa médica tradicional, existem inúmeras outras fontes com potencial para contribuir com o *big data*. Alguns exemplos incluem registros hospitalares, registros médicos de pacientes, resultados de exames médicos e muitos novos dispositivos de computação (conhecidos como "internet das coisas"), que são incorporados a objetos do cotidiano e podem coletar sinais vitais em tempo real de forma onipresente. Por meio da análise de grande quantidade de dados, doenças podem ser diagnosticadas mais precocemente, melhorando, assim, o prognóstico de doenças graves. Os custos médicos podem ser reduzidos, as epidemias podem ser previstas, as doenças podem ser evitadas e a qualidade de vida pode ser melhorada. A nível individual, caminhamos cada vez mais para a era da medicina personalizada, na qual as informações de um determinado paciente, especialmente os seus dados genéticos, serão analisadas e tratadas para o estabelecimento de um tratamento específico e personalizado.

Desde o início dos anos 2000, quando o termo *big data* entrou em uso, a forma como os dados são coletados e analisados mudou completamente. Os famosos 3 "Vs" de *big data* referem-se ao volume, velocidade e variedade.[2] Embora outras pessoas tenham acrescentado vários outros Vs a essa definição, como veracidade, valor, visualização e variabilidade, no final, estamos falando de dados com tamanhos que excedem a capacidade de processamento de softwares tradicionais dentro de um tempo e valor aceitáveis. Outra boa explicação para o *big data* é que ele "envolve todas as coleções de dados dotadas de um 'tamanho' e uma indefinição suficientes (tendo sido reunidas sem uma hipótese *a priori* ou uma tarefa de

[I]Médica assistente do Serviço de Cirurgia Torácica, Instituto do Câncer do Estado de São Paulo, Hospital das Clínicas HCFMUSP, Faculdade de Medicina, Universidade de Sao Paulo, Sao Paulo, SP, BR.
http://orcid.org/0000-0002-8378-7704

[II]MD, PhD. Professor titular do Programa de Cirurgia Torácica, Instituto do Coração, Hospital das Clínicas HCFMUSP, Faculdade de Medicina da Universidade de Sao Paulo, Sao Paulo, SP, BR. Diretor do Departamento Científico da Associação Paulista de Medicina, São Paulo (SP), Brasil.
https://orcid.org/0000-0001-7243-5343

*Este editorial foi publicado em inglês na revista São Paulo Medical Journal, volume 140, edição número 6, de novembro e dezembro de 2022.

pesquisa específica) para serem consideradas como territórios ainda amplamente intocados de onde se possam derivar novas percepções na forma de regularidades imprevistas".[3]

A Comissão Europeia desenvolveu a seguinte definição para *"big data in Health"*: "refere-se a grandes conjuntos de dados coletados rotineiramente ou automaticamente, que são capturados e armazenados eletronicamente. É reutilizável no sentido de dados polivalentes e compreende a fusão e ligação de bases de dados existentes com o objetivo de melhorar a saúde e o desempenho do sistema de saúde. Não se refere a dados coletados para um estudo específico".[4]

Na área médica, com os avanços da radiômica, por exemplo, método que extrai um grande número de características de imagens médicas por meio de algoritmos de caracterização de dados, milhões de dados podem ser extraídos automaticamente por meio de softwares desenvolvidos para essa finalidade, alimentando grandes repositórios de dados que serão usados para prever resultados, dispensar novos exames ou otimizar a investigação diagnóstica, reduzindo custos e melhorando o tratamento de inúmeras doenças.[5]

Imagine a seguinte situação: você, como profissional de saúde, tem um paciente com nódulo pulmonar detectado em tomografia computadorizada de tórax. Com uma única tomografia de tórax, por meio da análise da radiômica e da inteligência artificial, é possível determinar se o nódulo é maligno e qual o subtipo anatomopatológico, poupando o paciente dos riscos de uma biópsia. A análise do parênquima pulmonar, área cardíaca, calcificações coronarianas, massa muscular e tecido subcutâneo fornecerá informações sobre função pulmonar e cardíaca, sarcopenia e desnutrição, permitindo uma previsão precisa de complicações pós-operatórias e novamente evitando que o paciente seja submetido a inúmeros exames pré-operatórios. Além disso, uma análise detalhada das estruturas anatômicas desse pulmão, com o reconhecimento de possíveis variações anatômicas, permitirá um melhor planejamento cirúrgico, de forma que o cirurgião saiba de antemão as dificuldades que encontrará durante a cirurgia.

Este é o futuro e provavelmente será uma realidade em aproximadamente 10 anos. No entanto, para chegar lá, precisamos de grandes quantidades de dados, não apenas dados extraídos automaticamente de exames de imagem ou dispositivos médicos, mas também dados clínicos. Porque é assim que os sistemas aprendem, é assim que a inteligência artificial e o aprendizado de máquina funcionam. Isso significa que construir e alimentar bancos de dados clínicos é um passo essencial para chegar ao futuro do *big data*.

No entanto, construir e manter um banco de dados clínico não é uma tarefa fácil. Usando novamente o exemplo do câncer de pulmão, existem grandes bancos de dados multicêntricos internacionais[6,7,8] que ainda dependem do esforço humano para imputação de dados, apesar de já existir tecnologia para extração de informações diretamente dos prontuários eletrônicos ("Eletronic Health Records"- EHRs). Este fato causa uma grande dificuldade para a coleta de dados de qualidade em larga escala, especialmente em países como o Brasil, onde não há iniciativas nacionais ou governamentais para o desenvolvimento de registros médicos, e a grande maioria dos profissionais de saúde não está familiarizada com a coleta de dados fora da pesquisa clínica.

Existem muitos desafios associados ao *big data* na área da saúde. Mesmo nos Estados Unidos, onde a adoção de programas de EHR testados e certificados pelo governo federal no setor de saúde está quase completa, a existência de diferentes programas, com diferentes terminologias clínicas, especificações técnicas e capacidades funcionais têm levado a dificuldades na interoperabilidade e compartilhamento de dados.[2] Além disso, a maioria dos sistemas EHR contém muitos dados não estruturados, tornando mais complexa a extração de informações úteis para *big data*. Traçando um paralelo entre os Estados Unidos e o Brasil, nosso país está ainda mais distante dessa realidade, uma vez que muitos serviços de saúde ainda estão implementando seus sistemas de EHR ou simplesmente ainda utilizam prontuários manuais.

Além disso, ter um sistema de EHR disponível em um determinado serviço de saúde não significa que os dados serão realmente extraídos para contribuir com um banco de dados clínico. Isso porque, além de todas as dificuldades tecnológicas listadas acima, existem outras barreiras relacionadas a leis regulatórias sobre acesso e compartilhamento de informações pessoais, como a Lei Geral de Proteção de Dados ("LGPD") no Brasil, a Lei Geral de Dados Protection Regulation ("GDPR") na Europa e Health Insurance Portability and Accountability Act (HIPAA) nos Estados Unidos. Aprofundando nessa questão, ainda existem muitas dúvidas a serem discutidas na sociedade e no âmbito da pesquisa, especialmente no Brasil, onde a LGPD é relativamente recente. Em primeiro lugar, quem é o proprietário das informações dos pacientes - hospitais, pesquisadores ou os próprios pacientes? Assumindo que a informação pertence ao paciente, como isso pode contribuir para o desenvolvimento de *big data* e, finalmente, para a melhoria da própria saúde do paciente? Como dados sensíveis relacionados a questões de saúde podem ser coletados em grandes volumes com o consentimento dos pacientes e garantindo a proteção da privacidade?

Diante de tudo isso, apesar de todo o avanço da tecnologia, muitos dados de saúde ainda estão sendo "perdidos" diariamente, mesmo em grandes serviços de referência, simplesmente porque não há iniciativa para a coleta prospectiva desses dados. Enfrentar esse problema requer, antes de

tudo, reconhecer a importância da coleta de dados para o desenvolvimento da ciência em escala nacional. Precisamos urgentemente de iniciativas para o desenvolvimento de bancos de dados nacionais de especialidades médicas e registros de câncer. Precisamos conhecer nossos próprios dados para comparar nossos números com referências internacionais, ao invés de apenas consumir dados internacionais e tentar extrapolá-los para nossa realidade.

REFERÊNCIAS

1. Calude CS, Longo G. The Deluge of Spurious Correlations in Big Data. Found Sci. 2017;22:595-612. https://doi.org/10.1007/s10699-016-9489-4.
2. Dash S, Shakyawar SK, Sharma M, et al. Big data in healthcare: management, analysis and future prospects. J Big Data. 2019;6:54. https://doi.org/10.1186/s40537-019-0217-0.
3. Todde V, Giuliani A. Big Data. A briefing. Ann Ist Super Sanita. 2018;54(3):174-5. PMID: 30284542; https://doi.org/10.4415/ANN_18_03_02.
4. Study on Big Data in Public Health, Telemedicine and Healthcare. Luxembourg: Publications Office of the European Union; 2016. Disponível em: https://health.ec.europa.eu/system/files/2016-12/bigdata_report_en_0.pdf. Acessado em 2022 (19 ago).
5. Mayerhoefer ME, Materka A, Langs G, et al. Introduction to Radiomics. J Nucl Med. 2020;61(4):488-95. PMID: 32060219; https://doi.org/10.2967/jnumed.118.222893.
6. Falcoz PE, Brunelli A. The European general thoracic surgery database project. J Thorac Dis. 2014;6 Suppl 2(Suppl 2):S272-5. PMID: 24868445; https://doi.org/10.3978/j.issn.2072-1439.2014.04.20.
7. Goldstraw P, Crowley J. The International Association for the Study of Lung Cancer International Staging Project on Lung Cancer. J Thorac Oncol. 2006;1(4):281-6. https://doi.org/10.1016/S1556-0864(15)31581-1.
8. Sekine I, Shintani Y, Shukuya T, et al. A Japanese lung cancer registry study on demographics and treatment modalities in medically treated patients. Cancer Sci. 2020;111(5):1685-91. Erratum in: Cancer Sci. 2021;112(3):1332. PMID: 32103551; https://doi.org/10.1111/cas.14368.