Uso de *big data* em saúde no Brasil: perspectivas para um futuro próximo*

doi: 10.5123/S1679-49742015000200015

The use of big data in healthcare in Brazil: perspectives for the near future

Alexandre Dias Porto Chiavegatto Filho
Universidade de São Paulo, Faculdade de Saúde Pública, São Paulo-SP, Brasil

Resumo

O uso de *big data* tem crescido em todas as áreas da ciência nos últimos anos. Existem três áreas auspiciosas para o uso de *big data* em saúde: medicina de precisão (*precision medicine*); prontuários eletrônicos do paciente; e internet das coisas (*internet of things*). Entre as linguagens de programação mais utilizadas em *big data*, duas têm se destacado nos últimos anos: R e Python. Em relação às novas técnicas estatísticas, espera-se que técnicas de *machine learning* (principalmente as árvores de classificação e regressão), metodologias para controlar por associações espúrias (como a correção de Bonferroni e a taxa de falsas descobertas) e metodologias para a redução da dimensão dos dados (como a análise de componentes principais e o *propensity score matching*) sejam cada vez mais utilizadas. A questão da privacidade será também cada vez mais importante na análise de dados. O uso de *big data* na área da saúde trará importantes ganhos em termos de dinheiro, tempo e vidas e precisa ser ativamente defendido por cientistas de dados e epidemiologistas.

Palavras-Chave: Big Data; Metodologia; Estatística e Dados Numéricos; Brasil.

Abstract

The use of big data has increased in recent years in all scientific areas. There are currently three promising areas for the use of big data in healthcare: precision medicine, electronic medical records and the internet of things. Two programming languages have gained momentum in data science: R and Python. Regarding the statistical techniques, it is expected that machine learning (especially classification and regression trees), methodologies for controlling spurious associations (such as Bonferroni correction and false discovery rate) and methodologies for dimension reduction (such as principal components analysis and propensity scores) will be increasingly used. Privacy is an issue that will become ever more important in data analysis. The use of big data in healthcare will bring enormous gains in terms of costs, time and lives saved, and needs to be actively defended by data scientists and epidemiologists.

Key words: *Big Data; Methodology; Statistics and Numerical Data; Brazil.*

* Artigo baseado na disciplina "Introdução a *Big Data* em Saúde", ministrada pelo autor como curso de verão na Faculdade de Saúde Pública da Universidade de São Paulo (FSP-USP). Mais informações estão disponíveis em: http://www.fsp.usp.br/alexandre.

Introdução

Ao observar os avanços da ciência nos últimos anos, é possível encontrar fortes indícios de que a próxima grande fronteira da epidemiologia será a análise de grandes bancos de dados (*big data*). O crescimento do número de estudos multicêntricos e a pressão pela transparência dos gastos públicos têm aumentado a quantidade de dados disponíveis e criado uma demanda por novas formas de análise de dados complexos e desestruturados –um conjunto de técnicas conhecido como *data mining*.

Essa demanda por especialistas da área de *big data* pode trazer enormes oportunidades para os epidemiologistas, os profissionais com experiência em análise de dados em saúde. Caso os epidemiologistas acolham de braços abertos essa oportunidade, estarão em uma posição privilegiada para liderarem projetos de pesquisa em todas as áreas da saúde e dominarem o debate sobre as políticas públicas em saúde – principalmente em questões puramente quantitativas, como análises de custo-benefício e de impacto dos programas de saúde. No Brasil, algumas das oportunidades de análise de *big data* mais imediatas incluem o *linkage* dos bancos de dados mantidos pelo Ministério da Saúde, como o Sistema de Informações sobre Nascidos Vivos (Sinasc), o Sistema de Informações sobre Mortalidade (SIM), o cartão SUS, entre outros, além da colaboração entre centros de pesquisa nacionais e internacionais para o desenvolvimento de pesquisas multicêntricas.

> *Atualmente, define-se big data como uma quantidade de dados suficientemente grande que leve a uma mudança nas formas tradicionais de análise de dados.*

O primeiro desafio é definir o que é exatamente *big data*. Não se trata de um problema de solução fácil, já que a quantidade de dados usada pelas pesquisas aumenta a cada ano. Na metade do século passado, encontrar os parâmetros de uma regressão linear com 500 observações era uma tarefa que levava alguns dias. Hoje, são necessários também alguns dias para rodar modelos bayesianos com centenas de milhares de observações. Em vez de definir *big data* por meio de uma quantidade específica de *bytes*, ou pelo tempo necessário para a análise, uma melhor solução é enfatizar a necessidade de mudança de processos. Atualmente, define-se *big data* como uma quantidade de dados suficientemente grande que leve a uma mudança nas formas tradicionais de análise de dados.[1]

Novas áreas para a análise de *big data* em saúde

Apesar de a revolução do *big data* na saúde estar apenas começando, já é possível identificar três áreas auspiciosas para os próximos anos: a medicina de precisão, os prontuários eletrônicos do paciente e a internet das coisas.

Medicina de precisão

A maioria dos conhecimentos científicos ainda é baseada em grandes médias. Por exemplo, uma metanálise recente verificou que o uso de novos anticoagulantes orais diminui o risco de acidentes vasculares cerebrais (AVC) e eventos embólicos sistêmicos em 19%.[2] O problema aqui é que ninguém teve o risco diminuído em 19%. Algumas pessoas tiveram o risco diminuído em 100% (não tiveram um desses eventos) e as outras em 0% (tiveram um desses eventos).

Ou seja, sabemos apenas que o uso dos anticoagulantes orais diminui a presença dos eventos para a população como um todo – o resultado foi estatisticamente significativo com $p<0,0001$ –, mas não sabemos exatamente para quem. No referido estudo, os anticoagulantes orais não tiverem o efeito desejado para muitos pacientes: dos 29.312 indivíduos que receberam o medicamento, 911 tiverem um AVC ou evento embólico sistêmico.

Quem são as pessoas para as quais o medicamento não funciona? Talvez não funcione para mulheres com mais de 60 anos, com histórico de tabagismo, que tiveram pelo menos um filho, que têm uma mutação no gene G20210A e que moram em um bairro com concentração de material particulado inalável (MP_{10}) abaixo de 36 µg/m³. A verdade é que não sabemos.

A medicina de precisão (*precision medicine*) tem como objetivo ajudar a resolver esse problema. Em vez de prescrever o mesmo anticoagulante oral para todos os pacientes, espera-se que um dia seja possível indicá-lo apenas para indivíduos para os quais o medicamento verdadeiramente funcione. É claro que será muito difícil

atingir a precisão de 100%, devido à multicausalidade das doenças, mas se conseguirmos dobrar a eficácia atual de todas as intervenções de saúde o número de vidas salvas será inestimável.

Para que a medicina de precisão seja um dia de fato uma realidade, o mais importante será aumentar o tamanho das amostras das pesquisas. Isso será possível por meio de incentivos a novos estudos multicêntricos que usem a mesma metodologia e pelo *linkage* de dados públicos já existentes. A digitalização de todos os dados dos pacientes pelos serviços de saúde também será fundamental para estimular novas análises e aumentar o tamanho das amostras. De especial importância será a universalização do uso integrado do prontuário eletrônico do paciente.

Prontuário eletrônico do paciente

A realidade brasileira ainda é a dos prontuários específicos para cada unidade de saúde, digitalizados ou, em muitos casos, em papel. Assim como as prescrições em papel, os prontuários em papel dificultam a transferência, a atualização e a compreensão das informações. Além disso, o espaço físico necessário para o seu armazenamento tem gerado problemas logísticos aos sistemas de saúde e incentivado negativamente a introdução de novas informações.

Existe uma forte tendência para a universalização da digitalização dos prontuários no Brasil, principalmente nos grandes centros urbanos.[3] Apesar de necessária, essa novidade já chega defasada. A digitalização sem dúvidas traz mudanças positivas, mas o fato de esses prontuários não poderem ser acessados por profissionais de outros centros de saúde traz perdas de tempo, dinheiro e vidas.

Uma solução é o uso integrado do prontuário eletrônico do paciente (PEP), que permitiria o uso remoto do mesmo prontuário por todos os estabelecimentos de saúde. Alguns dos benefícios do uso integrado do PEP são o ganho de tempo no preenchimento, a diminuição do viés de memória/esquecimentos, a completitude das informações e o seu potencial para uso em pesquisas científicas.

O uso do PEP já é universal na atenção primária no Reino Unido, o que tem possibilitado um grande número de pesquisas científicas.[4] No caso do Brasil, cujo sistema de saúde possui uma atuação mais forte do sistema privado, a implantação dos PEPs será necessariamente mais complexa. Uma posição de liderança do SUS nessa questão será fundamental para garantir o uso integrado dos PEPs no futuro próximo.

Internet das coisas

Das três perspectivas para o uso de *big data*, a internet das coisas (*internet of things*) é no momento a realidade mais distante, apesar de alguns avanços recentes. A promessa é que um dia a maioria dos objetos de uso diário estará de alguma forma conectada à internet. Por exemplo, a geladeira, o chuveiro e até a porta das casas estarão conectados entre si pela internet. O sensor da porta poderá identificar quando o morador chega suando e informar automaticamente a geladeira, que prepara uma água gelada, e o chuveiro, que liga a água em uma temperatura morna.

As possibilidades de uso na área específica da saúde são imensas. No caso de idosos, por exemplo, se o chão da casa tiver um sensor conectado à internet, uma queda brusca de um corpo poderá gerar um alerta automático para os cuidadores do idoso e, em situações críticas, para o próprio sistema de saúde. Outra possiblidade promissora será o uso de *wearables*, objetos eletrônicos conectados ao corpo que poderão identificar a iminência de infartos e acidentes vasculares antes mesmo do próprio indivíduo.

A quantidade de dados gerados pela internet das coisas será imensamente útil aos epidemiologistas, já que permitirá identificar todos os passos imediatos e distantes que levaram ao aparecimento das doenças ou ao óbito. Enquanto atualmente ainda dependemos de pesquisas ativas, no futuro o desafio da ciência será convencer as pessoas a fornecerem os dados que já foram automaticamente coletados pela internet das coisas.

As várias linguagens de programação para análise de dados: a busca por um consenso

A grande quantidade de linguagens e *softwares* disponíveis para a análise de dados – Stata, SAS, SPSS, R, JMP, MATLAB, Julia, Python, entre outros – tem dificultado o compartilhamento de resultados e o desenvolvimento de novas análises. O uso de terminais para análise de dados por via remota em pesquisas multicêntricas tem significado que o *software* estatístico utilizado precisa ser o mesmo para todos os cientistas

que participam da pesquisa. A escolha do *software* de análise de dados passa então a ser fundamental e deve ser decidida por meio de um consenso entre todos os pesquisadores do grupo.

Duas linguagens de programação têm conquistado o apoio crescente dos cientistas na última década: R e Python.[5] A expectativa é que essas duas linguagens passem a ser dominantes também entre epidemiologistas. Ambas são *open source*, gratuitas e têm uma comunidade de programadores e cientistas extremamente ativa, o que significa que novas metodologias estatísticas são rapidamente incorporadas pelos usuários por meio de pacotes e bibliotecas.

A vantagem do Python é ser uma linguagem de programação geral, enquanto a base do R é a análise de dados. Originalmente o R era mais completo em relação ao número de metodologias estatísticas disponíveis, mas cada vez mais o Python tem se aproximado do R, principalmente graças ao crescimento do Pandas, uma biblioteca do Python especializada em análise de dados. A escolha entre aprender uma ou outra linguagem não é tão fácil, uma vez que depende dos objetivos do cientista, mas atualmente existe um movimento em direção a um consenso sobre o uso de uma dessas duas linguagens em análise de dados. É importante, entretanto, mencionar que a linguagem Julia tem crescido rapidamente nos últimos dois anos, o que torna o futuro do atual consenso incerto.

As duas linguagens têm se adaptado ao crescimento do uso de *big data*, com a introdução de pacotes específicos. No R, por exemplo, o pacote *big memory* permite o uso mais eficiente da memória RAM por meio da linguagem C++, e o pacote *ff* cria uma estrutura de dados que funciona como se estivesse na memória RAM, apesar de salvos no disco rígido. No Python, o *NumPy* permite o uso de matrizes multidimensionais na linguagem C, o que aumenta a velocidade da análise de dados. O fato de o R e o Python terem uma comunidade ativa de programadores é uma garantia de que serão desenvolvidas soluções para os novos problemas de *big data* no futuro.

Metodologias para *big data*

O crescimento da quantidade e complexidade dos dados tem também gerado alguns desafios em relação à escolha da metodologia estatística. As técnicas tradicionais de análise de dados apresentam algumas limitações para *big data*, principalmente em relação aos dados com muitas dimensões e no caso da presença de correlações espúrias. Além disso, o crescimento do número de profissionais de tecnologia da informação (TI) na área de análise de dados tem aumentado o interesse em *machine learning*.

Machine learning

Desde o início da Revolução Industrial, sempre se colocou a possibilidade de as máquinas chegarem um dia a ter iniciativa própria. Na área da análise de dados, isso significa a elaboração de algoritmos que respondam e se adaptem automaticamente aos dados sem a necessidade de intervenção humana contínua.

A metodologia de *machine learning* atualmente mais utilizada em análise de dados são as árvores de decisão (*decision trees*), que podem ser usadas quando a variável dependente assume valores finitos (árvore de classificação) ou valores contínuos (árvore de regressão). Analisemos aqui um exemplo bastante simples do uso de árvore de classificação, em que o objetivo é prever o fato de um município brasileiro ter um coeficiente de mortalidade infantil (CMI) abaixo da média nacional (14,7 óbitos para cada 1.000 nascidos vivos).

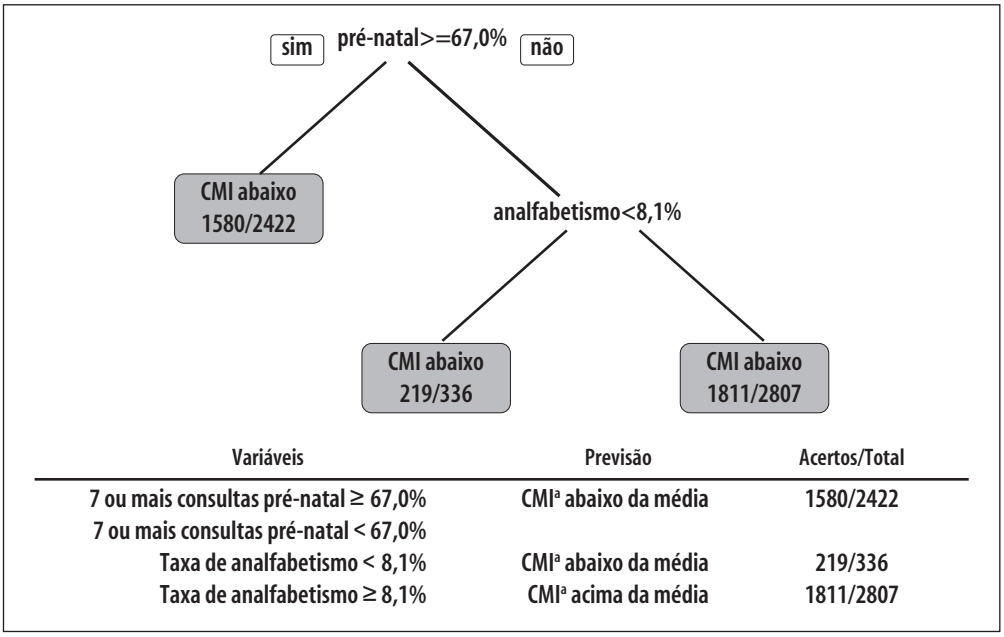
Como o CMI tem alta variabilidade anual em municípios pequenos, utilizou-se o período de 2008 a 2012 para garantir a estabilidade dos resultados. Por questão de simplicidade, serão incluídas apenas duas características dos municípios brasileiros: proporção de nascimentos com 7 ou mais consultas de pré-natal e taxa de analfabetismo, ambas referentes a 2010. Os dados das três variáveis para cada um dos 5.565 municípios brasileiros foram retirados do Datasus.[6]

Para a análise de árvore de regressão, foi utilizado o pacote *rpart* do R.[7] Os resultados podem ser replicados utilizando-se o código-fonte a seguir:

```
ML <- read.csv("https://sites.google.com/site/
alexandrechiave/mlexemplo/mlexemplo.csv")
CMI <- ML$CMI
CMI[CMI==0] <- "CMI abaixo"
CMI[CMI==1] <- "CMI acima"
prenatal<- ML$prenatal
analfabet<- ML$analfabet
install.packages ("rpart")
install.packages ("rpart.plot")
library ("rpart")
library("rpart.plot")
```

```
model.rpart <- rpart (CMI ~ prenatal + analfabet)
rpart.plot (model.rpart, type=0, extra=2, varlen=10)
png ("CMI.png")
rpart.plot (model.rpart, type=0, extra=2, varlen=10)
graphics.off()
```

Sem a necessidade de intervenção humana explícita, o algoritmo identificou dois pontos preditivos (conhecidos como os "nós" da árvore): 1) proporção de mulheres com 7 consultas de pré-natal acima ou igual a 67,0%; e 2) taxa de analfabetismo menor que 8,1%. O gráfico a seguir indica que o algoritmo, utilizando apenas duas variáveis, conseguiu identificar a posição correta dos municípios brasileiros em relação à média nacional em 64,9% dos casos (3.610 de 5.565).



Variáveis	Previsão	Acertos/Total
7 ou mais consultas pré-natal ≥ 67,0%	CMI[a] abaixo da média	1580/2422
7 ou mais consultas pré-natal < 67,0%		
Taxa de analfabetismo < 8,1%	CMI[a] abaixo da média	219/336
Taxa de analfabetismo ≥ 8,1%	CMI[a] acima da média	1811/2807

a) CMI = Coeficiente de mortalidade infantil

Gráfico 1 – Árvore de classificação para o coeficiente de mortalidade infantil, municípios brasileiros, 2008-2012

Trata-se obviamente de uma árvore de classificação extremamente simples, com apenas duas variáveis preditoras. Modelos mais elaborados já têm começado a aparecer em revistas internacionais da área da saúde. Uma análise recente, por exemplo, utilizou os dados do World Mental Health de 24 países para construir 20 grupos de risco para transtornos de estresse pós-traumático (TEPT) com técnicas de *machine learning*.[8] Apesar de a prevalência na amostra total ter sido de apenas 4,0%, no grupo de maior risco 56,3% dos indivíduos apresentaram TEPT.

As metodologias mais populares de *machine learning* apresentam várias limitações, principalmente o problema do sobreajuste e o possível aumento no número de associações espúrias (pois na maioria dos casos não há um embasamento teórico na formulação dos modelos), mas a expectativa para o futuro é que, com o seu acolhimento pela epidemiologia, esses problemas sejam confrontados e novas soluções apareçam.

Associações espúrias

A grande quantidade de variáveis utilizadas em *big data* tem como limitação o aumento do número de associações espúrias. O uso do valor de 0,05 para critério de significância é de certo modo adequado para uma única comparação. No caso de centenas ou milhares de testes de hipótese ao mesmo tempo, a possibilidade de uma associação aleatória ser considerada como significante é enorme. Uma solução simples é evitar a realização de testes de hipótese para todas as variáveis disponíveis, limitando-se apenas àquelas para as quais existe uma fundamentação teórica para a associação (em epidemiologia, isso significa jamais utilizar metodologias *stepwise*). O problema dessa estratégia é

que em alguns casos, como as análises epigenéticas, o embasamento teórico simplesmente ainda não existe.

Algumas metodologias têm sido historicamente utilizadas em pesquisas genéticas e epigenéticas para evitar associações espúrias, sendo as principais a correção de Bonferroni e a taxa de falsas descobertas. Com o crescimento do uso de *big data* na área de saúde, essas duas metodologias têm sido rapidamente incorporadas pela epidemiologia.

A correção de Bonferroni é a metodologia mais simples e mais tradicional para tentar diminuir o número de associações espúrias, com o objetivo de evitar a presença de um grande número de falsos positivos (ou seja, de rejeitar a hipótese nula quando ela é verdadeira). A metodologia estabelece um novo critério de significância, obtido pela divisão do valor original pelo número de hipóteses a serem testadas. No caso de serem feitos 8 testes de hipótese (h), o novo critério de significância será:

$$\alpha = \frac{\alpha^*}{h} = \frac{0{,}05}{8} = 0{,}00625$$

A correção de Bonferroni tem como limitação o fato de ser ultraconservadora, dado que fica cada vez mais improvável encontrar valores significativos à medida que o número de testes de hipótese aumenta. Uma metodologia alternativa, em uso crescente em genética e epigenética, é a taxa de falsas descobertas (*false discovery rate*, ou FDR). O objetivo nesse caso é controlar a proporção esperada de falsos positivos. Entre as hipóteses nulas rejeitadas pela pesquisa, a FDR é a proporção esperada de resultados falsos:

$$FDR = E\left[\frac{V}{V+S}\right]$$

Na equação, V é o número de falsos positivos e E o número de positivos verdadeiros (e, portanto, $V+S$ é o número total de resultados considerados significantes). Existem alguns métodos para se controlar a FDR, sendo os mais comuns as abordagens de Benjamini-Hochberg (BH) e a de Storey.[9] No primeiro caso, são calculados os p-valores para todos os testes de hipótese e ordenados de forma crescente. Para o caso de a FDR de interesse ser α, é necessário encontrar o maior k para o qual:

$$P_{(k)} \leq \frac{k}{h}\alpha$$

Na equação, h é o número de testes de hipótese testados. Assim, todos os testes de hipóteses (previamente ordenados) situados até k serão rejeitados, ou seja, serão considerados significantes. Matematicamente é possível provar que a FDR dessa análise será sempre menor do que α.

A abordagem de Storey é um pouco menos conservadora que a anterior, pois permite um maior número de rejeições da hipótese nula. É parecida com a BH, exceto pelo fato de que introduz no modelo a proporção das associações para as quais a hipótese nula é verdadeira (π_0), um valor que não é diretamente conhecido, mas que pode ser inferido utilizando-se diferentes técnicas:

$$\pi_0 P_{(k)} \leq \frac{k}{h}\alpha$$

Como π_0 é sempre igual ou menor do que 1, o número de testes de hipótese rejeitados será maior que no caso da BH. Quando π_0 for igual a 1, o resultado final será exatamente igual ao anterior.

Redução da dimensão dos dados

O objetivo da redução da dimensão dos dados é permitir uma melhor visualização dos dados, diminuir a quantidade de memória RAM necessária para rodar os modelos e diminuir a quantidade de associações espúrias. Duas metodologias estatísticas conhecidas em epidemiologia têm recebido atenção recente em *big data*: análise de componentes principais e *propensity score*.

O objetivo da análise de componentes principais é transformar muitas variáveis semelhantes em apenas alguns componentes principais linearmente não correlacionados. O primeiro componente principal será aquele que contém a maior quantidade de variação explicada pelos dados, seguido de outros componentes principais com menor. A quantidade de componentes principais selecionados dependerá do limite determinado pelo cientista para os autovalores (*eingenvalues*) da matriz de variância-covariância do modelo, normalmente igual a 1. O uso de componentes principais já é frequente em análises de *big data*, principalmente em estudos genéticos.

O *propensity score* também permite reduzir o número de variáveis, só que utilizando a probabilidade condicional de exposição (ou tratamento). É utilizado para garantir o balanceamento das variáveis entre os

expostos e não expostos, permitindo a comparação de indivíduos dos dois grupos com igual probabilidade de exposição. Formalmente é igual a:

$$e(x) = Pr(Z=1|x)$$

Na equação, Z é a exposição e x é o vetor de variáveis independentes, sendo calculado por meio de uma regressão logística. *Propensity scores* têm sido utilizados principalmente em estudos epidemiológicos,[10] mas o aumento do seu uso em outros estudos de *big data* tem sido defendido por outros pesquisadores.[11]

Muitos dados ≠ dados bons

Conseguir diferenciar entre a importância da quantidade e da qualidade dos dados não é tão simples. É possível identificar aqui três grupos: indivíduos sem conhecimento de estatística, indivíduos com um pouco de conhecimento de estatística e indivíduos que trabalham com estatística. O primeiro grupo costuma achar que a solução para todos os problemas das pesquisas científicas é aumentar o número de dados (costumam ser aqueles que acham que os erros das pesquisas de intenção de voto ocorrem por terem sido pesquisados apenas 2.000 eleitores). Indivíduos com pouco conhecimento de estatística têm uma visão exatamente oposta sobre *big data*: acham que a grande quantidade de dados torna a análise científica inválida devido a problemas de amostragem. O terceiro grupo já entende que lidar com amostras enviesadas sempre ocupou boa parte do tempo do cientista, mesmo antes da existência de *big data*, e que já existem algumas soluções para o problema e muitas novas aparecerão.

De fato, o uso de *big data* implica que em muitos casos as amostras disponíveis não serão representativas de toda a população. Por exemplo, dados de *smartphones* ou *wearables* serão provenientes majoritariamente de pessoas de alta renda e a adoção dos prontuários médicos dependerá do conhecimento tecnológico dos profissionais de saúde. Isso com certeza traz limitações às pesquisas, mas claramente também não as inviabiliza. Desde sempre, pesquisas epidemiológicas muito raramente utilizam uma amostra aleatória da população por questões de tempo e custo, sendo mais comum o uso de técnicas de amostragem, como estratificação e seleção de multiestágio.

Metodologias tradicionais já estão sendo incorporadas em *big data* para lidar com o problema de amostragem. A metodologia mais simples para controlar pelo viés da amostra é a adição de pesos de acordo com a representatividade de cada indivíduo em relação à população de interesse. Assim, indivíduos menos representativos terão peso e efeito menores nos resultados finais. Outra opção é realizar uma amostragem estratificada da própria amostra, considerando-se a distribuição das características da população de interesse. Como nos outros exemplos, o crescimento do uso de *big data* em pesquisas científicas deve acelerar o desenvolvimento de novas e mais complexas metodologias de amostragem.

Desafios para o futuro

Existe um consenso de que o grande desafio do uso de *big data* nos próximos anos será a questão da privacidade. O risco de uma grande quantidade de dados confidenciais ser roubado e divulgado será cada vez mais real. A solução para o problema será a conscientização dos cientistas sobre a importância da privacidade e o desenvolvimento de protocolos de segurança cada vez mais rígidos. Por exemplo, é cada vez mais comum a análise de dados exclusivamente dentro de um terminal de acesso restrito. Novas técnicas para garantir o sigilo dos dados, possivelmente utilizando técnicas de criptografia, serão cada vez mais incorporadas em pesquisas científicas.

Entretanto, a realidade é que certamente aparecerão escândalos de vazamento de dados sigilosos, seja por descuido de alguns cientistas ou por invasões propositais. Além de fazer de tudo para que isso jamais aconteça, é papel do cientista também informar a população sobre os imensos ganhos de tempo, dinheiro e vidas que a análise de *big data* traz para a sociedade. Esses escândalos, apesar de certamente prejudiciais para as vítimas, não podem ser utilizados para a restrição de pesquisas com *big data*.

A análise de *big data* encontra-se em um ponto de aceleração, que se tornou possível pela confluência de dois fatores: a pressão pela divulgação de resultados de pesquisas públicas e o desenvolvimento computacional necessário para as análises estatísticas. O potencial da análise de *big data* está apenas começando a virar uma realidade na área da saúde, e epidemiologistas estão na posição ideal para liderarem essa nova área. Apesar de existirem algumas limitações metodológicas e problemas de privacidade, a era do *big data* traz imensas oportunidades para o avanço do conhecimento em saúde.

Referências

1. Oxford English Dictionary [Internet]. Oxford: Oxford University Press; 2015. Big Data; [cited 2015 Apr 17]; [1 paragraph]. Available from: http://www.oed.com/view/Entry/18833
2. Ruff CT, Giugliano RP, Braunwald E, Hoffman EB, Deenadayalu N, Ezekowitz MD, et al. Comparison of the efficacy and safety of new oral anticoagulants with warfarin in patients with atrial fibrillation: a meta-analysis of randomised trials. Lancet. 2014 Mar;383(9921):955-62.
3. Câneo PK, Rondina JM. Prontuário eletrônico do paciente: conhecendo as experiências de sua implantação. J Health Inform. 2014 abr-jun;6(2):67-71.
4. Williams H, Spencer K, Sanders C, Lund D, Whitley EA, Kaye J, et al. Dynamic consent: a possible solution to improve patient confidence and trust in how electronic patient records are used in medical research. JMIR Med Inform. 2015 Jan-Mar;3(1):e3.
5. King J, Magoulas R. 2014 Data science salary survey: tools, trends, what pays (and what doesn't) for data professionals. Sebastopol: O'Reilly; 2014.
6. Ministério da Saúde (BR). Departamento de Informática do SUS. Informações de Saúde (BI) [Internet]. Brasília: Ministério da Saúde; 2015 [citado 2015 abr 14]. Disponível em: http://www2.datasus.gov.br/DATASUS/index.php?area=04
7. Varian HR. Big data: new tricks for econometrics. J Econ Perspect;28(2):3-28.
8. Kessler RC, Rose S, Koenen KC, Karam EG, Stang PE, Stein DJ, et al. How well can post-traumatic stress disorder be predicted from pre-trauma risk factors? An exploratory study in the WHO World Mental Health Surveys. World Psychiatry. 2014 Oct;13(3):265-74.
9. Storey JD. A direct approach to false discovery rates. J R Statist Soc B. 2002 Ago;64(3):479-98.
10. Chiavegatto Filho ADP, Kawachi I, Gotlieb SL. Propensity score matching approach to test the association of income inequality and mortality in São Paulo, Brazil. J Epidemiol Community Health. 2012 Jan;66(1):14-7.
11. Grimmer J. We are all social scientists now: how big data, machine learning, and causal inference work together. PS. 2015 Jan;48(1):80-3.