Machine Learning
Biopharma Applications and Overview of Key Steps for Successful Implementation

Mustafa Oguz, PhD
Senior Research Associate, Real-World Evidence, Evidera

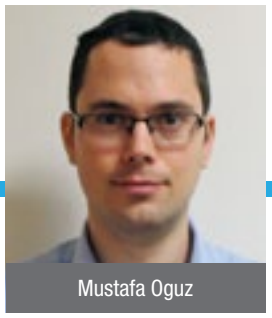
Andrew P. Cox, PhD
Research Scientist, Real-World Evidence, Evidera

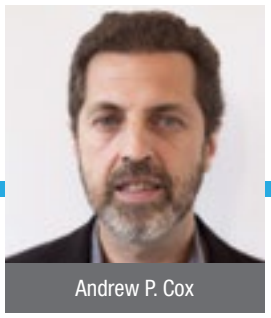
Introduction

Machine Learning (ML) is the science of programming computers to perform tasks based on rules learned from data instead of rules explicitly described by humans. Although statistical methods in health care for tasks such as stroke risk prediction[1] have been in use for a long time, three trends enabled the widespread adoption of ML applications in the past decade: increase in computing resources and cloud services that allow generation and storage of massive quantities of data; availability and digitization of diverse data sources (e.g., genomics databases, electronic health records, patient registries, large commercial databases, social media, and data collected through wearable technologies), and improvements in ML algorithms such as random forests, support vector machines, and deep learning, which can reveal complex relationships in data that simpler algorithms might miss.

Despite the advances in the adoption of ML methods in the pharmaceutical industry, there is room for increased application, especially in late stage development. According to a 2017 survey of 3,073 companies globally from 14 business sectors, only about 16% of health care firms adopted at least one artificial intelligence (AI) technology at scale or in a core part of their business, putting the health care sector behind high-tech and telecommunications (31%), finance (28%), and transportation (21%).[2] One reason behind the comparatively slow pace of adoption is a lack of clarity on the impact of AI methods on workflows in the pharmaceutical industry.[3]

In this article we review ML applications in the pharmaceutical industry that increase efficiency and allow more convincing value demonstration, broadly following a product's lifecycle from drug discovery to drug repositioning. Going from big data to improved efficiency in business and clinical benefits, however, requires at least a


Mustafa Oguz


Andrew P. Cox

broad understanding of the steps a project team needs to take to implement a successful data analysis project. In the second section, we describe these steps.

Applications of ML in the Pharmaceutical Industry

Drug Discovery

One of the most promising application areas for ML methods is new drug development, which is estimated to cost $2.6 billion on average.[4] Although computational methods have been employed for drug discovery for decades[5,6] (see Hiller, et al. for a 1972 study that applied artificial neural network in drug design), ML methods combined with large data sources enable access to deeper insights faster compared to traditional methods that mostly rely on numerous costly biochemical experiments.[7] For example, deep learning, an ML method based on discovering hidden layers of variables that connect the input data to outcomes, is used to predict drug-target interactions (DTIs)[8]; generate novel molecules predicted to be active against a given biological molecule[9]; predict cell-penetrating peptides for antisense delivery[10]; and, to model quantitative structure – activity and structure – property relationship (QSAR/QSPR) of small molecules to predict blood-brain barrier[11] permeability (see Ying Y, et al.[12] and Lo, et al. [5] for other ML applications on drug discovery).

Clinical Trial Site and Patient Selection

A study that analyzed data from 151 global clinical trials conducted by 12 companies at 15,965 sites found that 52% of clinical trials exceeded their planned enrollment timelines, with 48% taking significantly longer to complete enrollment.[13] Delays were more pronounced in clinical trials of the disease of the central nervous system, with an average planned timeline of 11 months vs. average actual timeline of 12.7 months. Companies also reported that on average 11% of sites in clinical trials failed to enroll any patients at all. ML algorithms can leverage historical data on site performance to maximize the probability that selected sites can deliver patients quickly, minimize drop-out rates, and adhere to the clinical protocol. ML models can be built using historical data on past performance, focusing on clinical trials, infrastructure, and time to first patient enrollment, which are predictive of future performance according to studies conducted by the industry.[14,15] Text mining of social media with natural language processing and predictive analytics applied to electronic health records are already being used by the industry to identify potential patients who might not have been formally diagnosed and who might be ideal candidates for recruitment into clinical trials for rare diseases.[16]

Wearables in Observational Studies and Clinical Trials

Increased miniaturization and longer battery life of electronics enabled the manufacturing of wearable devices that make collection of continuous and accurate medical data more practical than ever.[17] Wearable technologies include smartwatches, wristbands, hearing aids, electronic/optical tattoos, head-mounted displays, subcutaneous sensors, electronic footwear, and electronic textiles. ML methods are routinely employed to convert raw data collected from these technologies in observational studies and clinical trials to meaningful clinical end points. For example, Willetts et al. collected accelerometer data from 132 participants whose physical activities were labeled using video cameras to train ML models that can predict physical activity and sleep patterns.[18] The authors then used their results to label physical activity data collected from more than 96,000 UK Biobank participants. These algorithms can also be potentially used to classify patient data from clinical trials. A review of medical literature put the number of clinical trials that collected data from wearable devices as of late 2015 at 299.[19] An important disease area where biosensors can collect data that were previously unavailable to researchers is neurodegenerative diseases such as Alzheimer's disease. Biosensors worn by the patient and placed in the patient's home as part of a clinical trial can provide quantitative and continuous information on a subject's cognitive status and ability to perform daily tasks.[20]

> ML and natural language processing methods are commonly used to identify patient experiences related to treatments in the real world.

Pharmacovigilance

ML and natural language processing methods are commonly used to identify patient experiences related to treatments in the real world. Social media in general and patient forums in particular offer a rich source of information about adverse events and other problems associated with treatments. The U.S. Food and Drug Administration (FDA) encourages "external stakeholders to explore the use of social media tools such as medical community blogs, crowdsourcing, and social media pages" to identify patient perspectives regarding disease symptoms.[21] Social media content can be used to complement literature review findings, supplement focus groups, gather expert opinions, and elicit patient interviews. The FDA is also exploring the value of social media to inform occurrence of adverse events.[22] Extracting useful signals from large volumes of text data in social media is an active area of research. Recent examples include a study by Gupta and colleagues who used recurrent neural networks for semi-supervised learning of models to extract adverse event mentions from social media posts.[23]

Precision Medicine

Precision medicine is a prevention and treatment approach that considers a patient's genes, environment, and lifestyle.[24] According to a survey of 100 pharmaceutical

industry leaders, precision medicine has the potential to help accurately identify new drug targets; provide clarity regarding target patient profiles, thus, enabling more targeted clinical trials with smaller patient numbers and faster market access; reduce research and development (R&D) cycle length; and, more convincingly demonstrate benefits.[25]

Delivery of the premise of precision medicine depends on the ability to harmonize diverse data sources such as genomics, clinical trials, electronic health records, clinician notes, and wearables, and to develop predictive models to optimize treatment strategies. Recent studies on precision medicine emphasize methods to harmonize these different data sources. Rajkomar and colleagues[26] used deep learning methods to develop predictive models of mortality based on electronic health records and free text records from two hospitals; these models predicted the risk of inpatient mortality, unplanned readmission within 30 days, long lengths of stay, and discharge diagnosis. Recently Pai and Bader[27] reviewed ML algorithms that leverage patient similarity scores based on genomics data and electronic health records to identify subgroups of type 2 diabetes patients, predict tumour subtype in ependymoma, and predict treatment response.

Adherence Prediction

Non-adherence to medication is a major cause of revenue loss for the pharmaceutical industry and imposes a very high cost to public health care systems. A report on economic costs of medication non-adherence estimated the industry's annual revenue loss from non-adherence at $188 billion (or 37% of the $508 billion potential total revenue) in the U.S. alone and $564 billion globally.[28] The report further estimated that even a 10% increase in medication adherence across disease areas would increase the total annual revenue of the industry by $41 billion in the U.S. A systematic literature review in 2017 estimated that disease-specific, per patient, per year cost of non-adherence to medication ranges between $949 and $44,190 (in USD 2015).[29]

Predicting risk of non-adherence allows more targeted interventions to decrease non-adherence rates. Unsupervised ML methods can be used to identify non-adherent patient segments that display different characteristics and reasons for non-adherence to allow tailoring interventions to different patient groups. A recent example of non-adherence risk estimation includes a study by Krumme and colleagues[30] who used pharmacy and demographic predictors, pre-index adherence levels, and medical claims data to predict one-year adherence to statin treatments.

Drug Repositioning

Faced with growing R&D costs and low approval rates for new compounds, repositioning of existing drugs is a potential way to cut costs and expand to new indications. Drug repositioning has the benefit of reducing drug development time, since toxicity and safety profiles of drug candidates for repositioning have already been studied.[31] Before widespread use of systematic approaches and computational methods, such as similarity searching, text mining, and network analysis, drug repositioning was largely based on unexpected associations observed in clinical trials or in medical practice.[32] ML methods promise to accelerate this process. Examples include neural networks for prediction of sensitivity of cancer cells to drugs; support vector machines for prediction of drug therapeutic class; collaborative filtering and network analysis to predict drug-disease associations; and, text mining to leverage medical literature to highlight potential new indications for existing drugs.[33]

Implementing a Successful ML Project

Machine learning and artificial intelligence are written about and discussed extensively, in print and on websites, by a multitude of authors, including both companies and organizations involved in ML. The impression is often given that ML can be performed automatically in a "point and click" manner without particular specialist knowledge from analysts. Companies advertise services and packages that are able to apply ML and AI to problems in an automated manner. Whilst this may be true for certain specific applications like image classification, language translation, and other applications where no unmeasured variables are present and large volumes of data are available for pre-trained models, this is not the case for applications in the pharmaceutical and medical industries. Like other analytical approaches, such as that of traditional statistics, a detailed review and understanding of the problem and the data, as well as rigorous attention to methodological considerations is absolutely crucial. Inappropriate application of ML methods can lead to erroneous conclusions and inaccurate performance assessment. At worst, this can lead to mistakes in health care decisions which might be based on evidence derived from ML studies. Regardless of the ML application area, there are core steps in every ML project that must be followed to get actionable insights from data.

Building the Right Team

Building the right team or providing the core team with access to the required domain expertise is a stage of analytical projects that is often overlooked with potentially important consequences. ML has its origins in computer science with increases in computing power and availability of cloud computing making ML approaches to data analysis possible. Consequently, there are many cases where the team performing the ML analysis consists solely of computer scientists. Whilst individuals with a background solely in computer science are undoubtedly skilled in the application of ML, they often do not possess the skills or domain knowledge needed to apply the methods in the health sector. Whilst the emphasis of analysis in many sectors is often solely on predictive ability, this is not the

case in the health sector, where there is critical importance on inference, causality, rigor, understanding potential sources of confounding and bias, underlying epidemiology, and reasons why a particular method can be used for prediction. By failing to take into consideration these additional factors, critical errors can result. Similarly, analytic teams that consist of clinicians or epidemiologists may not apply the ML algorithms in a rigorous enough manner, leading to overfitting and resulting in over-optimistic prediction performance. It is, therefore, important to ensure that an analytic team consists of, or has access to, all the skillsets for a particular application, and even then, it is necessary that the multi-disciplinary team members are able to effectively communicate with each other.

Establish Whether ML is Necessary

Not all business questions that involve data analysis require an ML approach. The first question the project team needs to answer is whether it is possible to follow simple if-else rules to make predictions with enough accuracy. If so, then complicated algorithms might not be necessary. Another question is the availability of high quality and relevant data in enough quantity. Is there enough labeled data for the ML algorithm to learn from and, if not, how expensive is it to acquire more labeled data? Necessity and feasibility of ML approaches must be considered before committing more resources to an ML project. A phased approach, where a small feasibility study is conducted, can shed light on the decision to go ahead with an ML project or to prioritize quality data collection.

Formulate the Business Question as an ML Question

Despite the proliferation of ML algorithms, there is a limited number of ML tasks these algorithms can perform.[34] Formulating the business question in terms of one of these tasks is the first step towards a successful ML application. Different tasks include classification, regression, measuring similarity of entities, clustering similar groups together, identifying potential links between entities, data reduction, and causal modeling.[34] After the business question is cast as an appropriate ML task, the team must think hard about the metric that will be used to evaluate the model performance. In a classification task, for example, one pitfall is to simply look at the percentage of observations the model correctly classifies. This can be misleading in situations where even a simple decision rule (e.g., predict that no patient will experience the event of interest in the next year) would yield a high accuracy, simply because the event to be predicted is very rare. A model performance metric needs to incorporate the cost of different types of error (e.g., false negatives and false positives) especially if these have very high economic or health costs. Ideally the ML model must improve upon the methods currently in use as measured by the appropriate metric, whether those methods are based on expert judgment or existing risk scoring instruments.

Prepare Data for Analysis

According to a widely quoted estimate, data analysts spend 80% of their time collecting and preparing the data for analysis.[35,36] Because ML algorithms need quality data, and often in large quantities, data preparation is a very labor intensive part of any ML project. Activities at this stage include dealing with missing variables, creating new variables from existing ones that can boost model performance (feature engineering), and processing data so that it is usable by ML algorithms. All these steps require an understanding of the data sources, data fields, and subject matter knowledge. The team must consider how exactly the model will be used and which variables will be available to make new predictions when the model is deployed.

After the data is prepared for analysis, it is then necessary to randomly separate the dataset into a training validation set which will be used to train the models and assess their performance for purposes of model selection, and a test set to get an estimate of the selected model's performance when applied to data it has never seen before.

Train Models and Communicate Results

For any given ML task there is a large number of models from which to choose. Before going with the most complex model, such as a deep neural network with dozens of layers, it is better to start with simpler models such as logistic regression or random forests. Mean and standard deviation of different models' performance on validation sets can then be compared to select the best model. It is imperative to automate all these steps, including data preparation, because they involve extensive experimentation to find the right mix of features and models, as well as fine tuning the process. Note that model building and data preparation is an iterative process. Once model selection is complete, the team can use the test set to estimate the model's performance on new data. A model's parameters must never be tweaked to increase its performance on the test set. Otherwise, the real-world performance estimate will be biased.

When a model is selected, the analyst needs to go beyond reporting the model's performance and be able to answer the "so what" question from the business perspective, whether it relates to drug discovery or identifying undiagnosed patients. Assumptions, methods, and other technical details should be clearly laid out for more technical audiences.

Maintain the Model

A model's performance depends on whether new observations to which it is applied have characteristics similar to observations on which it was trained. It is likely that over time the characteristics of patients, clinical sites, or the instances it is being asked to make predictions for will change, leading to erosion of the model's accuracy. To prevent this decline, a model's parameters must be

tuned as new data is available to make sure that the initial performance is maintained or improved upon. Repeating model training with new data instances is therefore usually necessary. A related problem is application of a model to a new setting. A model trained on data from one region, patient population, or disease area is unlikely to perform as well when applied to another.

Conclusions

Whilst adoption of ML in early stage development has been widespread, use in later stage development is relatively early in its evolutionary path. Use cases for ML are still being developed and understood. There is no doubt that ML approaches can yield benefits in terms of efficiencies, new insights, and actionable evidence. However, knowledge of appropriate use of ML methods and potential applications is not widespread in our industry, a factor which is likely to slow its adoption. Whilst there may be something of a misconception that ML can be "automated" and can produce almost "magical" results with little effort, this is not the case. ML is an analytical technique and is best thought of in the same manner as traditional statistical analysis. It requires a range of specialist knowledge and rigorous application with attention to detail in the medical and pharmaceutical industries. ■

For more information, please contact Mustafa.Oguz@evidera.com or Andrew.Cox@evidera.com.

REFERENCES

1. Gage BF, Waterman AD, Shannon W, Boechler M, Rich MW, Radford MJ. Validation of Clinical Classification Schemes for Predicting Stroke: Results from the National Registry of Atrial Fibrillation. *JAMA.* 2001 Jun 13;285(22):2864-70.
2. Bughin J, Hazan E, Ramaswamy S, Chui M, Allas T, Dahlstrom P, Henke N, Trench M. Artificial Intelligence: The Next Digital Frontier? McKinsey Global Institute. Discussion Paper. June 2017. Available at: https://www.mckinsey.com/~/media/McKinsey/Industries/Advanced%20Electronics/Our%20Insights/How%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/MGI-Artificial-Intelligence-Discussion-paper.ashx. Accessed September 19, 2018.
3. Wilson CJ. Pharma's Path to Adopting AI and Other Emerging Technologies. Pharma R&D Today. 2018 Feb 7. Available at: https://pharma.elsevier.com/pharma-rd/pharmas-path-adopting-ai-emerging-technologies/. Accessed September 19, 2018.
4. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. *J Health Econ.* 2016 May; 47:20-33. doi: 10.1016/j.jhealeco.2016.01.012. Epub 2016 Feb 12.
5. Lo YC, Rensi SE, Torng W, Altman RB. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov Today.* 2018 Aug;23(8):1538-1546. doi: 10.1016/j.drudis.2018.05.010. Epub 2018 May 8.
6. Hiller SA, Golender VE, Rosenblit AB, Rastrigin LA, Glaz AB. Cybernetic Methods of Drug Design. I. Statement of the Problem - The Perceptron Approach. *Comput Biomed Res.* 1973 Oct;6(5):411-21.
7. Wang Q, Feng Y, Huang J, Wang T, Cheng G. A Novel Framework for the Identification of Drug Target Proteins: Combining Stacked Auto-Encoders with a Biased Support Vector Machine. *PLoS One.* 2017 Apr 28;12(4):e0176486. doi: 10.1371/journal.pone.0176486. eCollection 2017.
8. Lee I, Nam H. Identification of Drug-Target Interaction by a Random Walk with Restart Method on an Interactome Network. *BMC Bioinformatics.* 2018 Jun 13;19(Suppl 8):208. doi: 10.1186/s12859-018-2199-x.
9. Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular De-Novo Design through Deep Reinforcement Learning. *J Cheminform.* 2017 Sep 4;9(1):48. doi: 10.1186/s13321-017-0235-x.
10. Wolfe JM, Fadzen CM, Choo ZN, Holden RL, Yao M, Hanson GJ, Pentelute BL. Machine Learning To Predict Cell-Penetrating Peptides for Antisense Delivery. *ACS Cent Sci.* 2018 Apr 25;4(4):512-520. doi: 10.1021/acscentsci.8b00098. Epub 2018 Apr 5.
11. Wang Z, Yang H, Wu Z, Wang T, Li W, Tang Y, Liu G. In Silico Prediction of Blood-Brain Barrier Permeability of Compounds by Machine Learning and Resampling Methods. *ChemMedChem.* 2018 Aug 15. doi: 10.1002/cmdc.201800533. [Epub ahead of print]
12. Jing Y, Bian Y, Hu Z, Wang L, Xie XS. Deep Learning for Drug Design: An Artificial Intelligence Paradigm for Drug Discovery in the Big Data Era. *AAPS J.* 2018 Mar 30;20(3):58. doi: 10.1208/s12248-018-0210-0.
13. Lamberti MJ, Mathias A, Myles JE, Howe D, Getz K. Evaluating the Impact of Patient Recruitment and Retention Practices. *Ther Innov Regul Sci.* 2012 July 13; 46(5):573-580. doi: 10.1177/0092861512453040.
14. Getz KA. Predicting Successful Site Performance. Applied Clinical Trials. 2011 Nov 01. Available at: http://www.appliedclinicaltrialsonline.com/predicting-successful-site-performance. Accessed September 19, 2018.
15. Yang E, O'Donovan C, Phillips J, Atkinson L, Ghosh K, Agrafiotis DK. Quantifying and Visualizing Site Performance in Clinical Trials. *Contemp Clin Trials Commun.* 2018 Jan 31;9:108-114. doi: 10.1016/j.conctc.2018.01.005. eCollection 2018 Mar.
16. Salzman S. Rare Disease Recruitment Models Evolving: The Impact of Real-World Outcomes and Trial Design in the Rare Disease Arena. *The CenterWatch Monthly.* 2018 Feb;25(2). Available at: https://www.trinetx.com/wp-content/uploads/2018/02/cwm2502_FeatureReprint_TriNetX.pdf. Accessed September 19, 2018.

17. Yetisen AK, Martinez-Hurtado JL, Ünal B, Khademhosseini A, Butt H. Wearables in Medicine. *Adv Mater.* 2018 Jun 11:e1706910. doi: 10.1002/adma.201706910. [Epub ahead of print]

18. Willetts M, Hollowell S, Aslett L, Holmes C, Doherty A. Statistical Machine Learning of Sleep and Physical Activity Phenotypes from Sensor Data in 96,220 UK Biobank Participants. *Sci Rep.* 2018 May 21;8(1):7961. doi: 10.1038/s41598-018-26174-1.

19. Ricci M. Realising the True Potential of Health Wearables. *Pharmaphorum.* 2018 Sep 19. Available at: https://pharmaphorum.com/views-and-analysis/realising-true-potential-health-wearables/. Accessed September 19, 2018.

20. Teipel S, König A, Hoey J, Kaye J, Krüger F, Robillard JM, Kirste T, Babiloni C. Use of Nonintrusive Sensor-Based Information and Communication Technology for Real-World Evidence for Clinical Trials in Dementia. *Alzheimers Dement.* 2018 Sep;14(9):1216-1231. doi: 10.1016/j.jalz.2018.05.003. Epub 2018 Jun 21.

21. U.S. Food and Drug Administration (FDA). Patient-Focused Drug Development: Collecting Comprehensive and Representative Input. Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders. Draft Guidance. 2018 June. Available at: https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM610442.pdf. Accessed September 19, 2018.

22. U.S. Food and Drug Administration (FDA). Data Mining at FDA. Available at: https://www.fda.gov/ScienceResearch/DataMiningatFDA/default.htm. Accessed September 19, 2018.

23. Gupta S, Pawar S, Ramrakhiyani N, Palshikar GK, Varma V. Semi-Supervised Recurrent Neural Network for Adverse Drug Reaction Mention Extraction. *BMC Bioinformatics.* 2018 Jun 13;19(Suppl 8):212. doi: 10.1186/s12859-018-2192-4.

24. U.S. National Library of Medicine (NIH). What Is Precision Medicine? Available at: https://ghr.nlm.nih.gov/primer/precisionmedicine/definition. Accessed September 19, 2018.

25. Danner S, Solbach T, Ludwig M. Capitalizing on Precision Medicine: How Pharmaceutical Firms Can Shape the Future of Healthcare. *Strategy&.* 2017 Aug 24. Available at: https://www.strategyand.pwc.com/reports/capitalizing-precision-medicine. Accessed September 19, 2018.

26. Rajkomar A, Oren E, et al. Scalable and Accurate Deep Learning with Electronic Health Records. *npj | Digital Medicine.* 2018 May 8. Available at: https://www.nature.com/articles/s41746-018-0029-1. Accessed September 19, 2018.

27. Pai S, Bader GD. Patient Similarity Networks for Precision Medicine. *J Mol Biol.* 2018 Sep 14;430(18 Pt A):2924-2938. doi: 10.1016/j.jmb.2018.05.037. Epub 2018 Jun 1.

28. Forissier T, Firlik K. Estimated Annual Pharmaceutical Revenue Loss Due to Medication Non-Adherence. Capgemini Consulting. 2012 Nov. Available at: https://www.capgemini.com/wp-content/uploads/2017/07/Estimated_Annual_Pharmaceutical_Revenue_Loss_Due_to_Medication_Non-Adherence.pdf. Accessed September 19, 2018.

29. Cutler RL, Fernandez-Llimos F, Frommer M, Benrimoj C, Garcia-Cardenas V. Economic Impact of Medication Non-Adherence by Disease Groups: A Systematic Review. *BMJ Open.* 2018 Jan 21;8(1):e016982. doi: 10.1136/bmjopen-2017-016982.

30. Krumme AA, Franklin JM, Isaman DL, Matlin OS, Tong AY, Spettell CM, Brennan TA, Shrank WH, Choudhry NK. Predicting 1-Year Statin Adherence Among Prevalent Users: A Retrospective Cohort Study. *J Manag Care Spec Pharm.* 2017 Apr;23(4):494-502. doi: 10.18553/jmcp.2017.23.4.494.

31. Papapetropoulos A, Szabo C. Inventing New Therapies Without Reinventing the Wheel: The Power of Drug Repurposing. *Br J Pharmacol.* 2018 Jan;175(2):165-167. doi: 10.1111/bph.14081.

32. Bolgár B, Arany Á, Temesi G, Balogh B, Antal P, Mátyus P. Drug Repositioning for Treatment of Movement Disorders: From Serendipity to Rational Discovery Strategies. *Curr Top Med Chem.* 2013;13(18):2337-63.

33. Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z. A Survey of Current Trends in Computational Drug Repositioning. *Brief Bioinform.* 2016 Jan;17(1):2-12. doi: 10.1093/bib/bbv020. Epub 2015 Mar 31.

34. Provost F, Fawcett T. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media. 2013 August.

35. KDnuggets. CrowdFlower 2016 Data Science Report. Available at: https://www.kdnuggets.com/2016/04/crowdflower-2016-data-science-repost.html. Accessed September 18, 2018.

36. Lohr S. For Big-Data Scientists, "Janitor Work" is Key Hurdle to Insights. *The New York Times.* 2014 Aug 17. Available at: https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html. Accessed September 19, 2018.