ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

Machine Learning in the biopharma industry

Thibault Helleputte[1], Gaël de Lannoy[2] and Paul Smyth[2]

1- DNAlytics
Chemin du Cyclotron 6, 1348 Louvain-la-Neuve - Belgium

2- GSK Vaccines - Research & Development
Rue de l'institut 89, 1330 Rixensart

Abstract. Modern high-throughput technologies deployed in R&D for new health products have opened the door to Machine Learning applications that allow the automation of tasks and support for data-driven risk-based decision making. Appealing opportunities of applying Machine Learning appear for the development of modern complex drugs, for biomanufacturing production lines optimization, or even for elaborating product portfolio strategies. Nevertheless, many practical challenges make it difficult to apply Machine Learning models in the biopharmaceutical field. Innovative approaches must thus be considered in many of these practical cases. This tutorial paper is an attempt to describe the landscape of Machine Learning application to the biopharmaceutical industry along three dimensions: opportunities, specificities or constraints and methods.

1 Introduction

Machine learning in the health and pharmaceutical areas have attracted a great deal of attention over recent years (see e.g. [1] for a deeper review), including much publicised work from some of the major players in the area, such as Google DeepMind's work on the protein folding problem [2]. The aim of this short tutorial is to take a step back from the high profile, headline grabbing work to assess the opportunities and difficulties for machine learning in pharmaceutical industry. Traditional pharmaceutical products mostly involve small chemical molecules (e.g. pain killers), bio-technologies involve products based on larger molecules, such as proteins (products are then generally called biologics), but also (modified) viruses (products are then prophylactic vaccines) or even more recently gene and cell therapies. Healthcare in general also encompass medical devices (MD) and in vitro diagnostics (IVD), implants and much more. This paper focuses only on therapeutic and prophylactic products, at the exception of companion diagnostics (CDx): these IVDs are specifically used to predict the response of a particular therapeutic or prophylactic product, and are thus closely related to it. In some cases the product and the CDx are even designed together.

The biopharmaceutical industry relies heavily on advanced science and technology, but at the same time is quite conservative. This is due to the simple fact that biopharmaceutical products go into to human bodies and this implies a heavy regulatory burden to ensure the safety of patients.

Without entering into too many details the classical lifecycle of health products starts with product discovery, then follows with production process development, clinical validation, market access (obtain authorization to sell the product, and often also funding from health organizations or governmental agencies to pay for these products), production, and maybe ultimately merging and acquisition (which, finally, is driven by portfolio strategies). Most therapeutics are patented but after about 20 years, a loss of protection for the intellectual property occurs. At that stage, the industry often starts the development of generics (pharma) and biosimilars (biotech).

In this tutorial, we list a series of opportunities that arise in the different stages of health products lifecycle described above in Section (2). Section (3) points then at the multiple specificities of the field and the constraints they generate on the potential application of a Machine Learning (ML) approach towards the listed opportunities. Finally, Section (4) goes over the ML methods that are typically used to address these opportunities, and how their choice, or the way they are implemented (for example with respect to their metaparameter or their proper evaluation) is impacted or shaped by the identified constraints. It is of course not possible to be exhaustive in any of these sections.

2 Opportunities

2.1 Drug discovery

Everything starts with the search for a new product. This field is heavily relying on high-throughput molecular biology data (DNA sequencing, RNA sequencing, imaging, proteomics etc.). A central challenge in this field is the prediction of 3D structure and function of a molecule [2], generally larger molecules such as proteins, based on a simpler 1D or 2D structure. Another challenge is to acquire knowledge about a nucleic acid sequence. ML can be used in this case to predict sites of a nucleic acid strand corresponding to splicing sites or regulatory regions [3].

On a different note, literature about medical and biological findings has become far too large to be digested by human researchers. Natural language processing (NLP) of large domain specific corpora is thus a track of much interest to speed up research for new drugs.

2.2 Production and Process development

During production and development of a given product, a large selection of heterogeneous data types are collected: descriptors of raw materials, descriptors of the equipment used and the operations performed, QC laboratory measurements, environmental monitoring data (microbiology a.o.), values in time-series from sensors in bioreactors, fermentors or freeze-dryers, imaging of vials or syringes from the fill & finish phases and images of cultures from microscopy devices. Multiple perspectives of usage of these data exist. ML can be used in a yield maximisation or quality maximization perspective, to make predictions about sources of contamination or to minimize down-time and drive preventive maintenance. A major area of opportunity here is in ML-driven automation of redundant, noisy and time-consuming tasks. One such example is counting viral foci or bacterial colony forming units in pre-clinical and technical development, and in environmental monitoring during production. See Naets et al and Beznik et al in this session.

ML models and feature selection algorithms are also now being retrospectively applied on batch process data, in order to identify features correlated to the yield and/or to the quality and stability of the product. The insight gained on the process can then lead to process improvement and understanding.

Modern process development now also heavily relies on Raman optic probes and other spectroscopy devices like near-infrared technologies generating a spectra. Those techniques allow to quantify a very large amount of metabolites in one shot of analysis and almost in real-time. However, in order to turn the spectra into quantitative mg/ml measurements, blind source separation algorithms (to denoise the spectrum) and supervised ML algorithms (correlating the peaks in the spectra to amounts of known compounds) must be applied.

2.3 Clinical development

Clinical development is also an active playground for ML applications in multiple scenarios. Feature selection and prediction of response or adverse events can be used jointly to better understand the mechanisms of action of a product.

Using a multivariate model trained for a specific disease diagnosis, when the diagnosis is complex via traditional means, can help obtaining more homogeneous clinical trial cohort [5].

With respect to operations, predictive modeling of disease progression for example can be used to screen patients before their actual enrolment into a clinical trial. In this kind of strategy, the global study population gets enriched with strong or fast progressors (for example), which might positively impact the global study cost, duration or cohort size [6]. Also with respect to operations, patients recruitment curves can be predicted based on ML approaches, in order to support cost estimates, and refine schedule of trial completion or tailor an investigational drug supply chain to actual needs [7].

2.4 Market access

Obviously, for return on investment reasons, drug manufacturers ideally prefer to get an indication as wide as possible for their new drug, i.e. it is better if the drug is prescribed and funded for all individuals suffering from a given disease. However, there is in some cases a need for targeting a specific sub-population suffering from the disease instead. Main reason for such a choice is a lack of efficacy observed on some patients during clinical validation. A similar reasoning holds in case of adverse events suspected to occur only in a subpopulation of patients. Whatever the reason, identifying biomarkers of response/adverse events is a typical case where feature selection and predictive modeling are weapons of choice. Their results will constitute the basis for what is called companion diagnostics. Biomarkers research will typically be based on clinical descriptors of patients, (molecular) biology covariates, and drug safety/efficacy profiles.

Obtaining market access will also involve proving the benefit/risk ratio of a new product which is the object of Health Technology Assessment (HTA) which methodological aspects have been thoroughly described [4]. For therapeutic products without CDx, the application of this methodology does not require specific ML tasks, even if it requires data. However, when a CDx comes into play, the setup of a sensitivity/specificity threshold for that CDx has also to be determined. The Incremental Cost-Effectiveness Ration (ICER), which is central to HTA, can be used as an objective function in a ML setting directly to optimize a predictive model for the CDx.

2.5 Portfolio strategies

The biopharma industry has typically two models to develop a new drug: either develop it in-house using internal R&D expertise, or to acquire a process and a product developed elsewhere for scaling it up. To predict what drug to develop/acquire next, or predict which candidate drug is likely to achieve success in clinical trials can also be the subject of predictive modeling [8].

3 Specificities and constraints

The use cases listed in the previous section are fit for applications of ML techniques, but is subject to a series of field-specific constraints. We list below a series of them.

3.1 Dimensionality

ML relies heavily on three components: algorithms, computing power, and data. Algorithms are relatively mature, and more and more "off-the-shelf" implementations are available. Computing power has become a commodity via the numerous cloud computing infrastructure providers. But at all stages of the healthcare product lifecycle, data represent an issue: the ratio between number of features and number of samples is most of the time very high. In clinical studies, patients are rare and/or costly to recruit. In manufacturing, even large multinational will only produce a few tens of product batches per year. At the same time, more and more features are generated on each of them. The normal setting in the field is thus "fat data" rather than "big data". Having a 1000:1 ratio or more is really frequent. Traditional statistics are not fit for such a setting, and most ML approaches still make assumptions on statistical concepts that are thus here difficult to estimate.

3.2 Heterogeneity

There are many ways to look at the living. Most technology platform capture only one angle: DNA, RNA, proteins, imaging... To correctly grasp the concepts at hand (a drug mechanism of action, or the compatibility between donors and receivers in cell therapy for example) will require multiple angles of view. This will have as a consequence to generate data from heterogenous nature. Either because of very different numerical ranges and format (images, spectrum, text, audio, ...), or shapes of distributions (lognormal, normal, exponential), or even information nature (binary, categorical, continuous, times-series, graph-based). Not all ML techniques are capable of dealing with all these specificities, and the assumptions they make might get violated by one of these characteristics.

3.3 Noise

The data in the field are affected by noise. This noise can be in the features themselves. A spectrometer will partially be impacted by dirts or substrate, a sample will have less well tolerated snap drying. If this error is more or less randomly distributed, then increasing the number of samples analyzed will allow to overcome the issue (but we said above that the field frequently lacks samples in sufficient number). In omics data, sequencing platforms make reading errors. The additional issue here is that these errors are not (for most platforms) randomly distributed. As a result, increasing sample size will not remove the issue, and it is thus very likely that an ML algorithm will detect a technology switch rather than an actual signal of interest. The intent here is to stress the importance of prospective experiment planning, as an integral par of ML, specially in this field.

The noise can also be in the labels. For example image segmentation is based on manual drawings of segmentation boundaries by experts. It is not rare that even when working carefully, the boundaries are not perfectly drawn. As a result, some voxels corresponding to bone for example might fall just outside of the segmented area. This implies that the basis of the training of ML models will be containing errors. The same appears when some biological or imaging readouts are interpreted by a handful of experts: they do not always agree. So to some extent, it is irrelevant in some cases to go above a certain value of the performance metrics (because going above would imply to learn the errors in the labels), and some other considerations might be more useful (clinical performance, ...).

3.4 Regulatory limitations

In principle, a great opportunity of ML is the possibility of designing feedback loops, making a model continuously learn with experience, namely from its error. In the healthcare sector, there is almost an impossibility to benefit from ML feedback loops in routine, for regulatory constraints. Let's take the example of CDx which decision would rely on a ML model. To make the product available on the market, one has to register the device or IVD with an announced performance rate. These performances can simply not change from day to day. Similarly, in a production environment, if a method for screening product defaults by imaging is implemented and validated, this method cannot change from day to day.

Still, what is feasible is to accumulate data, and to update the model once in a while, after having controlled its updated performance (and the associated documentation).

Privacy is also limiting (for a good reason), the extent to which data in different care institutions can be accessed and combined. More advanced distributed learning approaches might then be considered [13]. These techniques are also fit to address pre-competitive collaboration frameworks.

3.5 Others

There are yet many other limitations and challenges related to ML in biopharma industry, such as:

- Largely unbalanced datasets.
- Biases appearing from batch of data to batch of data, making them in the best cases logically comparable but numerically very different, sometimes not comparable at all without specific approaches.
- Intelligibility of model predictions might be an issue for the general acceptance of this kind of approach. We can argue on the fact that this is fair (example: most pain killers MoA are not understood, yet, they sell by billions), but we can't purely reject this argument.

4 Machine Learning Methods

In this section, we provide a non-exhaustive list of methodologies that help addressing the opportunities but also taking into account the constraints above.

- Supervised classification algorithms such as deep learning, support vector machines, nearest neighbor systems, decision trees and random forests, partial least-square regressions. These algorithms must however be largely adapted in order to reduce the impact of label noise [12] and of class unbalance [9] that biomedical applications typically face.
- Blind source separation algorithms such as principal components analysis and independent component analysis, often optimized for time series or spectral data prefiltering [11];
- Feature selection algorithms, either embedded into the classifier (i.e. through regularization of parameters [14] or importance metrics [17]) or as a prior step (i.e. correlation, ANOVA, etc.) [10]; These feature selection methods can also be used in a context of transfer learning [15] or prior knowledge incorporation [16].
- Image processing algorithms such as U-Net and other variants improved to adapt to bioprocess images such as microscope pictures of bacterial colony forming units (see papers in the session from Beznik & al., Naets & al.).

5 Acknowledgments

The development of this article was sponsored by GlaxoSmithKline Biologicals SA. Gael de Lannoy and Paul Smyth were employees of the GSK group of companies at the time of the development of this article, and report ownership of shares and/or restricted shares in GSK.

References

[1] Ching, Travers, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface* 15.141 (2018): 20170387.

[2] Senior, A.W., Evans, R., Jumper, J. et al. Improved protein structure prediction using potentials from deep learning. *Nature* 577, 706â710 (2020). https://doi.org/10.1038/s41586-019-1923-7

[3] M. Pertea , X. Lin , S. L. Salzberg . GeneSplicer : a new computational method for splice site prediction . *Nucleic Acids Res* . 2001 Mar 1;29(5):1185-90 .

[4] L. Annemans. *Health-Economics for non-economists.* Pelckmans Pro. 2018.

[5] Olesen, TK, Denys, M-A, Goessaert, A-S, et al. Development of a multivariate prediction model for nocturia, based on urinary tract etiologies. *Int J Clin Pract.* 2019; 73:e13306. https://doi.org/10.1111/ijcp.13306

[6] T. Helleputte, Y. Henrotin, More efficient DMOAD trials through innovative screening strategies, *Annals of Rheumatic Diseases*, 78 (2019).

[7] Abdelkafi, C., Beck, B.H.L., David, B., Druck, C. Balancing Risk and Costs to Optimize the Clinical Supply Chain - A Step Beyond Simulation. *J Pharm Innov* 4, 96-106 (2009). https://doi.org/10.1007/s12247-009-9063-5

[8] Malas, T.B., Vlietstra, W.J., Kudrin, R. et al. Drug prioritization using the semantic properties of a knowledge graph. Sci Rep 9, 6281 (2019). https://doi.org/10.1038/s41598-019-42806-6

[9] G. de Lannoy, D. Francois, J. Delbeke and M. Verleysen. Weighted SVMs and Feature Relevance Assessment in Supervised Heart Beat Classification. Communications in Computer and Information Science, 2011, Volume 127, pp. 212-225.

[10] G. Doquire, G. de Lannoy, D. Francois, M. Verleysen. Feature selection for supervised inter-patient heart beat classification. Computational Intelligence and Neuroscience, 2011, No. 643816, pp. 1-9

[11] G. de Lannoy, T. Castecalde, J. Marin, M. Verleysen, J. Delbeke. Elimination of electrocardiogram contamination from vagus nerve recordings using ICA. Proceedings of the 14th Annual International FES Society Conference, 2009, pp 109-111.

[12] B. Frenay, G. de Lannoy and M. Verleysen. Label Noise-Tolerant Hidden Markov Models for Segmentation: Application to ECGs. Proceedings of the 2011 European Conference on Machine Learning, 2011.

[13] Jochems, A. et al. Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital â A real life proof of concept. *Radiotherapy and Oncology*, Volume 121, Issue 3, 459 - 467

[14] Abeel T., Helleputte T., Van de Peer Y. and Saeys Y., Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, *Bioinformatics*, Volume 26, No. 3, pp. 392-398. (2010)

[15] Helleputte T. and Dupont P. Feature Selection by Transfer Learning with Linear Regularized Models, *European Conference on Machine Learning (ECML)*, Bled, Slovenia, September 7-11, 2009.

[16] Helleputte T. and Dupont P. Partially Supervised Feature Selection with Regularized Linear Models, 26th *International Conference on Machine Learning (ICML)*, Montreal, Canada, June 14-18, 2009.

[17] J. Paul and P. Dupont, Inferring statistically significant features from random forests, *Neurocomputing* 2015.