Assessing Drug Development Risk Using Big Data and Machine Learning

Vangelis Vergetis[1], Dimitrios Skaltsas[1], Vassilis G. Gorgoulis[2,3,4,5], and Aristotelis Tsirigos[6,7]

ABSTRACT

Identifying new drug targets and developing safe and effective drugs is both challenging and risky. Furthermore, characterizing drug development risk, the probability that a drug will eventually receive regulatory approval, has been notoriously hard given the complexities of drug biology and clinical trials. This inherent risk is often misunderstood and mischaracterized, leading to inefficient allocation of resources and, as a result, an overall reduction in R&D productivity. Here we argue that the recent resurgence of Machine Learning in combination with the availability of data can provide a more accurate and unbiased estimate of drug development risk.

Introduction

With less than 10% of potential drugs that start clinical development eventually entering the market (1, 2), drug development is a very risky endeavor. Biotechnology and pharmaceutical companies are well aware of this challenge as they spend an average of more than 10 years and invest north of $2.5B for each drug that gets approved by regulators (3). Within this context, the ability to accurately assess the risk of drug development, or, on the flip side, the probability of technical and regulatory success (PTRS) of a particular development program is critical. We define PTRS as the probability that a development program will receive approval by the regulator (For example, the FDA in the US), i.e., the lower the PTRS, the higher the risk.

First, this ability to accurately assess the PTRS can allow drug development companies to prioritize among the different programs in their pipeline, make better resource allocation decisions, and ultimately increase the productivity of R&D efforts and investment (4). Which development program should progress to the next phase of clinical development (e.g., from phase I to phase II), and which one should not? What is the overall risk profile of the pipeline? Is the company taking too much risk, or is it playing it too safe? What can be done to de-risk a particular development program? These are all questions that can be answered by a comprehensive and unbiased model that accurately predicts the PTRS of specific programs.

Second, it can provide a broader view of risk and optimize resource allocation, as companies can also estimate the risk associated with external molecules that they can potentially acquire/in-license. This allows them to better value the potential of those molecules and/or to also make comparisons with their own development programs in the same indication. For clarity, we define "development program" as the combination of a specific molecule (or combination of molecules) and a particular indication, in a given phase of development. For example, a development program can be a phase II clinical trial that aims to evaluate the combination of a PD-1 molecule and a CTLA-4 molecule, for the treatment of melanoma. As a result, a particular drug (say a PD-1 inhibitor) might result in several different development programs – for example, one for say, breast cancer in phase 1, another one for melanoma in phase 3, etc.

Current Approach

In broad terms, the current approach that the industry uses to estimate the PTRS of a development program is based on:

i. Historical estimates driven by (i) the current state of the program (beginning of phase I, end of phase II, etc.), and (ii) the specific disease (breast cancer, relapsing-remitting multiple sclerosis, hypertension, etc.).
ii. Expert input from experienced physicians and drug developers, known in the industry as Key Opinion Leaders (KOL).
iii. Statistical (typically both univariate and multivariate) analyses performed by the R&D analytics groups within pharma companies and biotechnology companies. These analyses (e.g., refs. 5, 6) typically take into account several parameters (e.g., data from phase I, the mechanism of action of the drug, the availability of patients for the trial, etc.).

The typical approach is to start with an initial estimate based on (i) above–say, for example, 44% for a drug in breast cancer that has just started its phase III trial (1). This number is then adjusted based on qualitative input from KOLs (e.g., depending on beliefs around the specific biology of the drug, the specific phase I and phase II safety and efficacy data, etc.), and input from the analytics team of the sponsor company.

To our knowledge, there have only been early/nascent efforts to utilize Machine Learning (ML) in the process described above. On the basis of our own experience, we see several reasons for this

[1]Intelligencia Inc., New York, New York. [2]Molecular Carcinogenesis Group, Department of Histology and Embryology, Faculty of Medicine, School of Health Sciences, National Kapodistrian University of Athens, Athens, Greece. [3]Biomedical Research Foundation, Academy of Athens, Athens, Greece. [4]Molecular and Clinical Cancer Sciences, Manchester Cancer Research Centre, Manchester Academic Health Sciences Centre, University of Manchester, Manchester, United Kingdom. [5]Center for New Biotechnologies and Precision Medicine, Medical School, National and Kapodistrian University of Athens, Athens, Greece. [6]Institute for Computational Medicine, New York University School of Medicine, New York, New York. [7]Department of Pathology, New York University School of Medicine, New York, New York.

Corresponding Author: Aristotelis Tsirigos, NYU Langone Medical Center, 435 East 30th Street, New York, NY 10016. Phone: 646-501-2693; E-mail: Aristotelis.Tsirigos@nyulangone.org

Cancer Res 2021;81:816–9

doi: 10.1158/0008-5472.CAN-20-0866

underwhelming use of ML in drug development to-date. Some of those reasons are technical, and some of those reasons are cultural.

First, pharma companies seek to primarily utilize their own data in their statistical analyses, and not a larger industry-wide dataset that is more representative of successes and failures of development programs. Second, and as we describe in more detail in the section below, efforts often start with ambition but get quickly abandoned as practitioners realize that the required effort to put together the appropriate data set for ML algorithms to be trained on is significantly higher than initially expected. At the end of this article we discuss importance of data quality and completeness and show that it contributes significantly to an increase in overall predictive power.

In addition, there are also several cultural reasons behind this slower adoption of ML in drug development. Pharmaceutical companies are naturally focused on developing molecules, and up until very recently, executives had little patience for ML-driven efforts that did not produce positive results within a few weeks or months. It unfortunately takes a much more sustained commitment to build the right data infrastructure and models from scratch, and, as a result, early efforts have been abandoned. Furthermore, most pharmaceutical companies have well-defined processes, which create some inertia if new approaches are not properly embedded within the existing decision-making fabric of the organization. In other words, technology is often half the battle; the other half is getting the technology accepted and widely deployed within companies that have been used to a particular approach for several years. This said, we have noticed some cultural shifts lately, with several pharmaceutical and biotechnology executives becoming much more attuned to the potential upside from the use of ML models, and therefore being more patient as their investments in the space mature. In our own view, the trend has started, but there is still a lot of progress to be made before digital/ML approaches are fully embraced.

Within this context, we describe below an approach that is based on (i) a comprehensive dataset of successes and failures of development programs across the industry and (ii) ML-driven models, instead of the statistical approaches described above.

The Era of Machine Learning

Machine Learning, a subset of techniques within the broader umbrella of Artificial Intelligence (AI), is becoming increasingly utilized in the pharmaceutical industry and in the broader biomedical sciences field. To more precisely define the goals of AI, promote its fair and ethical use, and develop a broader vision for AI in this space, specialized industry groups are being formed, for example, the Alliance for Artificial Intelligence in Healthcare (AAIH). It is not our intent here to comprehensively describe those efforts (see ref. 7 for a more comprehensive review), but they indeed span different areas: from drug discovery, to efforts around protocol design, clinical trial execution (e.g., patient recruitment, site selection, biomarkers, etc.), diagnostics and imaging (e.g., classifying histopathology images), precision medicine/personalized treatment (e.g., matching existing drugs to patients based on molecular data), and risk assessment. This short article focuses on the latter: The use of machine learning to estimate PTRS.

Overall approach

There have been a few efforts that focus on predicting the success of clinical development programs (see, for example, refs. 8, 9). Some are based on more traditional/statistical approaches, while others utilize more advanced machine learning techniques. Here, we propose a systematic approach that incorporates more than 100 different factors across five broad areas:

i. *Clinical trial design*: Choice of primary/secondary endpoints, number of arms, inclusion/exclusion criteria, type of comparator used, use of biomarkers, number of patients, number of sites, etc.
ii. *Clinical trial outcomes*: The reported outcomes of the drug in previous studies. For example, the published phase II data for a drug that is currently transitioning to phase III, etc.
iii. *Regulatory data*: Any signals/designations by regulatory agencies. For example, prior approvals in other indications/Therapeutic Areas, breakthrough therapy designation, accelerated approval pathway, etc.
iv. *Drug biology*: Mechanism of action, modality, genetic and epigenetic alterations, target gene expression, tumor immunogenicity (for immuno-oncology drugs), molecular structure, etc.
v. *Sponsor characteristics:* Experience of the company running the development program in the specific disease area, etc.

Once the data described above are collected and curated for all development programs that have been both approved as well as failed in the past ~20 years (i.e., thousands of both successful and unsuccessful programs), a machine learning model can be trained to identify patterns in all that data and accurately predict the PTRS of the development program of interest.

We also note here that some of the feature categories above are specific to therapeutic areas and/or specific indications. For example, within (ii) above, oncology trials measure different outcomes, for example, objective response rate (ORR), overall survival (OS), progression-free survival (PFS), than trials in rheumatoid arthritis (e.g., American College of Rheumatology 20/50/70 criteria, etc.) or Parkinson's [e.g., Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS/UPDRS), ON/OFF, etc.]. Similarly, the most relevant drug biology features under (iv) above can differ from disease area to disease area. As a result, we argue – and indeed our own experience confirms this – that while overarching machine learning models can be trained across therapeutic areas, models that are specifically tailored to an indication (e.g., Parkinson's) or a family of indications (e.g., solid tumors) are likely to perform better.

Data challenges

Before we go into the performance of the proposed model, it is important to emphasize some of the practical challenges involved in building such a model.

First, there are dozens of data sources that need to be accessed and harmonized. This includes a significant software engineering challenge to build the appropriate data pipelines that are regularly updated, as well as address issues with data integration and consistency. Data can come from multiple sources, for example, (i) readily available public sources (e.g., clinicaltrials.gov, TCGA, ICGC, KEGG and REACTOME pathways, etc.); (ii) data in the private domain that typically require a subscription (e.g., data providers like Informa, full-text publications, etc.); (iii) data in the public domain that needs to be curated (e.g., conference abstracts, press releases, scientific articles, etc.); and (iv) private/in-house data owned by different pharmaceutical companies (e.g., assay data, patient data, drug metabolism and pharmacokinetics, etc.) and/or real-world data (e.g., for precision medicine, claims data) owned by healthcare providers and insurance companies. The format, conventions, and even ontologies (e.g., classification of drug

modalities, gene targets, drug names, etc.) that are used are typically different across data sources, and they need to be unified in a consistent way. In addition, one needs to keep track of when the data is released, especially clinical trial outcomes data: improperly incorporating such data without taking into account the release dates could result in models that suffer from forward-looking biases (see more details below).

Second, and perhaps more importantly, the need for human curation and quality control, even in the era of "big data" and AI, should not be underestimated. Data is typically missing, mistakes in data coding are prevalent, and the list of difficulties goes on. Furthermore, the data with the most predictive power is oftentimes not neatly arranged in a database but needs to be put together by content experts. For example, the clinical trial outcomes data in (2) above is dispersed across thousands of journal publications, press releases, conference abstracts, etc. Well-trained biologists and MDs would need to review this literature and capture the relevant data using a consistent methodology. Given the complexity of this task, Natural Language Processing methods can potentially assist the experts and cut down the required time for identifying the most relevant data (for example, by training an NLP model to identify the most relevant publications and then highlight the relevant sections within those publications), but, unfortunately, this process would be extremely hard to fully automate.

In our experience, more than 80% of the effort can typically go in data-related tasks described above, leaving not more than 20% of the effort towards training the different machine learning models.

Explainability

A final point on the challenges of using machine learning in areas like risk assessment is the need for transparency and "explainability." Drug developers need to know more than just the PTRS of their development program. In order to both develop some confidence around this prediction, as well as be able to potentially influence it – for example, improve the PTRS by making changes to the clinical trial protocol (see ref. 9) – they need to know the features that contribute to the overall PTRS estimate. This is not an area where a "black box" approach is likely to work.

As an example, consider that although it is definitely helpful to know that the estimated PTRS of a phase II program in breast cancer is, say, 26% (above historic average), it is even more informative to know what drives that probability up or down. Is it the mechanism of action of the drug or the choice of endpoints in the clinical trial? Or, is this based solely on positive initial data (for example ORR)? And to what extent is this prediction influenced by the fact that the sponsor company has limited experience in the space, and that the drug has not received any positive signals (e.g., breakthrough therapy) from the FDA? As a result, some popular machine learning approaches (e.g., deep learning) are less applicable in this particular domain, particularly given the importance of making the link between the output of the model and decision-making in real life.

Performance of Machine Learning Models

Within the context described above, we trained several machine learning algorithms – random forests, k-nearest neighbors, Support Vector Machines (SVMs), logistic regression, Gradient Boosting Trees, etc. The purpose here is not to go deep into the ML training methodology that was used, and readers who seek more detail should refer to our Technical Supplement (10). But we should nonetheless point out that we used the standard practices that are broadly acceptable in the AI community: we were careful to avoid forward-looking biases by only using relevant features, imputed missing data and binned certain continuous variables as needed, avoided data leakage by processing any data within the cross-validation folds, applied nested cross-validation to select the optimal models and used bootstrapping/bagging to assess their variance and any potential overfitting. Performance metrics are reported on unseen test sets.

The models were rigorously tested under different conditions, for example:

- Different therapeutic areas and indications (e.g., Oncology, Inflammation, Immunology, Central Nervous System diseases, etc.);
- Different stages of development (e.g., phase Ia/b, phase II, phase III);
- Different types of molecules (e.g., first-in-class, versus targets that have been previously validated/approved, etc.); and
- Different timeframes (e.g., prior to 2010, 2010–2017, post 2017).

Overall performance

We will not provide the full assessment in this short paper, but we will illustrate some average results. Using the standard AUC (area under the ROC curve) metric for a randomly chosen test set, our models achieve a performance of 0.81–0.93 depending on the different scenarios mentioned above.

To illustrate one specific example, when testing the algorithm across several hundred randomly chosen solid tumor programs at the beginning of phase II, the AUC is 0.89. Put another way, the algorithm correctly predicts more than 77% of the development programs that eventually receive regulatory approval, and more than 85% of the programs that do not. This performance is intriguing, particularly given how early those predictions are made (development programs at the beginning of phase II can be 4–6 years away from a regulatory decision), and how complex the overall problem is.

Perhaps the most important question is whether this model can generalize and, as a result, predict the future – in other words, prospective testing. To explore this scenario, we trained several models using data from development programs that either succeeded or failed in the 2000–2016 timeframe, and then tested the performance of the model on development programs in the 2017–2019 timeframe. To use the same example of early phase II solid tumor programs above, the model achieved an AUC of 0.90. Put differently, the algorithm correctly predicted more than 80% of the development programs that eventually received regulatory approval, and more than 85% of the drugs that didn't.

As a word of caution, we emphasize that we do not advocate that drug development decisions should rely solely on the recommendations of machine learning models. Drug development is an incredibly complex process, and there is still "art" involved in it. But we do argue that drug development experts in Portfolio Review Committees and/or Business Development decision-makers can utilize these machine learning models to calibrate and potentially remove any biases from their decision. These models do not rely on just a subset of successful or unsuccessful development programs. They have uncovered complex patterns across all such programs and can offer an input that will allow drug developers to make more informed decisions and with greater confidence.

Comparison to more traditional approaches

Further to the evaluation described above, we also compared the performance of our Machine Learning methodology to the performance of two "traditional" approaches:

i. Regression models, both univariate and multivariate, that take into account the most predictive or the three most predictive features respectively. Comparing those models across different assumptions (e.g., phases of development) we note that the AUC of our ML methodology yields significantly higher AUC (by 0.05–0.20) compared with these more simplistic models. For example, our ML methodology achieves an AUC of 0.91 when it assesses programs at the start of Phase 1, while the univariate model achieves an AUC of 0.71 and the multivariate model an AUC of 0.77. Please see (10) for more details.
ii. A Machine Learning model that is trained on a limited set of only publicly available data. Comparing the two models across different assumptions we note that the AUC of our ML methodology is significantly higher than the AUC of this more simplistic model (by 0.15–0.33). For example, our ML methodology achieves an AUC of 0.88 when it assesses programs at the end of Phase 2, while the ML model trained on the publicly available data achieves an AUC of 0.73. Please see (10) for further details.

One interesting conjecture from this analysis is that, for this particular problem, the quality and completeness of the underlying dataset is perhaps more important than the actual ML methodology that is used. Of course, both contribute to an increase in AUC and our methodology attempts to improve on previous efforts on both of those fronts, but when it comes to the relative importance of those two elements, data quality and completeness ranks higher. We would of course not want to generalize, and significant more work is needed before once can make any conclusive statements. Nevertheless, both our experience in the sector and the analysis above suggest that data availability is perhaps more important than the finetuning of Machine Learning models and parameters.

Conclusion

There is inherent risk in drug development. But if we utilize advances in AI coupled with curated industry-wide data to understand it and quantify it better, then allocation of resources can be much more effective, waste can be minimized, important/lifesaving drugs can enter the market faster, and patients can be better served. Simply put, two things are at stake: hundreds of millions of dollars, and a substantial positive impact on the lives of millions of patients – not necessarily in that order.

It is also clear that to fully capture the potential of applying ML in the broader drug development space, more work needs to be done that goes above and beyond the technical aspects of training state-of-the-art algorithms. For example, it is crucial for companies to create well annotated databases that bring together a multitude of different data in order to support this approach. Further, ontologies (for example around drug modalities, technologies, platforms etc.) need to be carefully defined in order for those databases to have a common "lexicon" and be able to seamlessly connect to one another. Finally, we would argue for greater transparency and data sharing between pharmaceutical companies. We believe that within the bounds of confidentiality and proprietary information, careful sharing of aggregated data can lead to an overall improvement in program success rates across for all the different companies in the industry.

Authors' Disclosures

V. Vergetis reports an executive role at Intelligencia Inc. outside the submitted work; in addition, Dr Vergetis has a patent for US 2020/0321083 pending and a patent for WO 2020/172131 pending. D. Skaltsas reports an executive role at Intelligencia Inc. outside the submitted work; in addition, Dr Skaltsas has a patent for US 2020/0321083 pending and a patent for WO 2020/172131 pending. A. Tsirigos reports personal fees from Intelligencia during the conduct of the study and personal fees from Intelligencia outside the submitted work. No disclosures were reported by the other authors.

Acknowledgments

The authors would like to acknowledge the numerous contributions by Gerasimos Liaropoulos, Maria Georganaki, Andreas Dimakakos, and by the broader team at Intelligencia Inc. Also, for feedback and advice on earlier parts of this work, we acknowledge David Baggett, Alexia Iasonos, Xhenete Lekperic, Rajneesh Nath, Aiman Shalabi, and Amit Singhal. Intelligencia is funded by venture capital funding and private investors. This work is funded partially by the National Public Investment Program of the Ministry of Development and Investment/General Secretariat for Research and Technology in the context of: (i) the application of Artificial Intelligence in the development of pharmaceuticals and innovative therapies (GSRT/2020ΣM01300002, awarded to V.G Gorgoulis), and, (ii) the Flagship Initiative to address SARS-CoV-2 (GSRT/2020ΣE01300001, awarded to V.G Gorgoulis).

Received March 17, 2020; revised October 14, 2020; accepted December 14, 2020; published first December 22, 2020.

References

1. Thomas DW, Burns J, Audette J, Carroll A, Dow-Hygelund C, Hay M. Clinical development success rates 2006–2015. BIO, Biomedtracker, and Amplion 2016. Available at https://www.bio.org/press-release/bio-releases-largest-study-ever-clinical-development-success-rates.
2. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol 2014;32:40–51.
3. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R&D costs. J Health Econ 2016;47:20–33.
4. Aitken M, Kleinrock M, Nass D, Simorellis A. The changing landscape of research and development: innovation, drivers of change, and evolution of clinical trial productivity. IQVIA Institute Report, April 2019. Available at https://www.iqvia.com/insights/the-iqvia-institute/reports/the-changing-landscape-of-research-and-development.
5. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. Contemp Clin Trials Commun 2018;11:156–164.
6. Jardim DL, Groves ES, Breitfeld PP, Kurzrock R.. Factors associated with failure of oncology drugs in late-stage clinical development: A systematic review. Cancer Treat Rev 2017;52:12–21.
7. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and Development. Nat Rev Drug Disco 2019;18:463–77.
8. Lo AW, Siah KW, Wong CH. Machine learning with statistical imputation for predicting drug approvals. Revised May 2019. Available at https://hdsr.mitpress.mit.edu/pub/ct67j043/release/9.
9. Getz KA, Campo RA. Trends in clinical trial design complexity. Nat Rev Drug Discov 2017;16:307.
10. Vergetis V, Liaropoulos G, Georganaki M, Dimakakos A, Skaltsas D, Gorgoulis VG, et al. A Machine Learning approach for assessing drug development risk. bioRxiv 2020.10.08.331926. DOI: https://doi.org/10.1101/2020.10.08.331926.