Igor Kruglyak
Senior Advisor, Avenga

Michael DePalma
Founder and President, Pensare LLC

Improving the risk-reward calculus for clinical trials

How natural language processing and machine learning can boost success in drug development

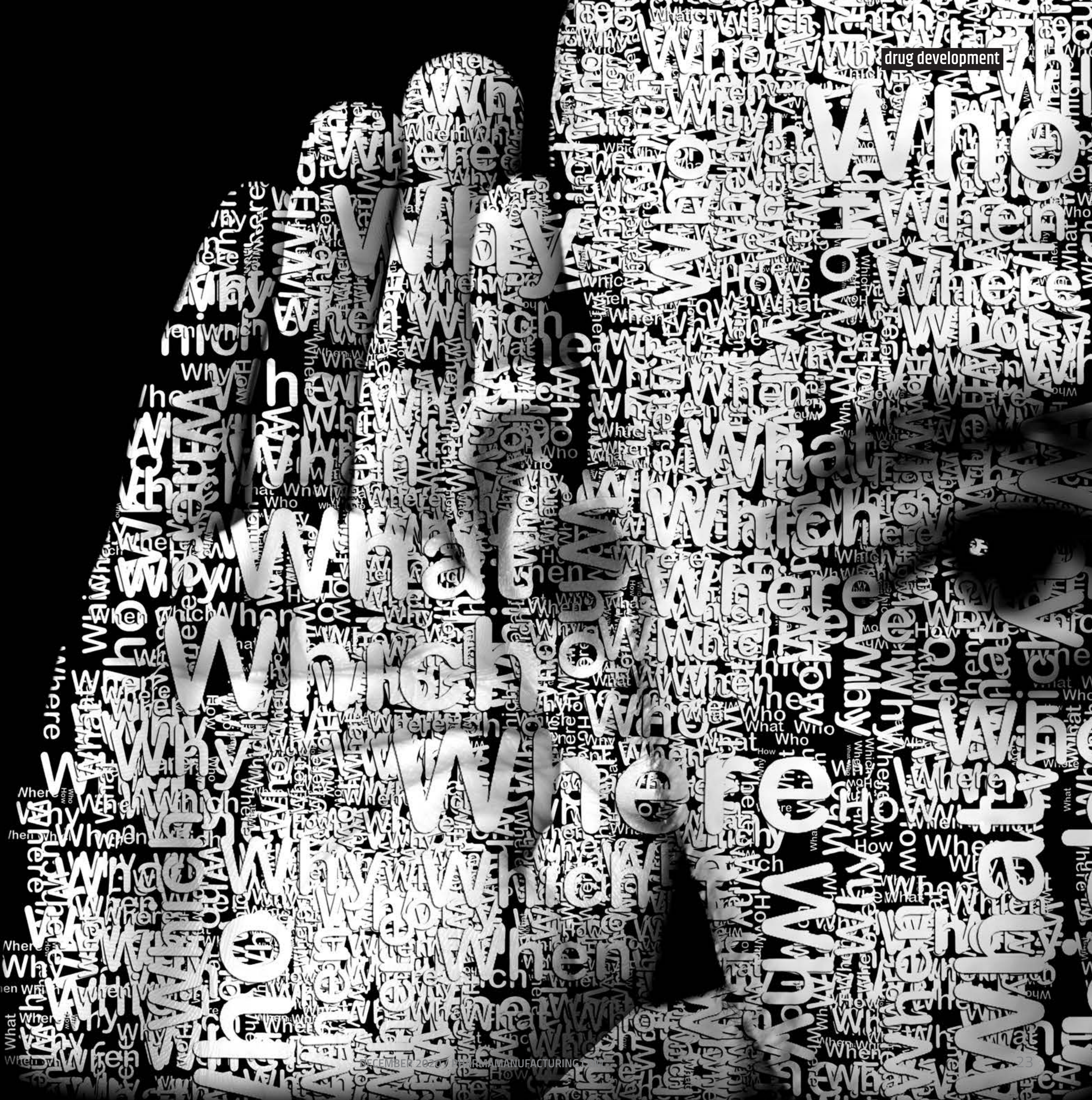
Bringing a new drug to the market is a monumental undertaking in terms of risk, resources and time. Typically, it involves budgeting $2 billion and an average of 10-12 years. A huge amount of these funds and some five to seven years are spent on clinical trials. Yet the risk-reward calculus for clinical trials remains dicey.

Despite an ever-increasing knowledge base and great technological advancements, a high percentage of clinical trials still fail. The reasons for this are manifold, ranging from questions of safety and efficacy to tightened regulations, to lack of funding and low patient recruitment rates.

According to an analysis published in *Nature* last year, the chances of success for a compound entering trials is under 10 percent. This result is falling in line with other estimates over the last couple of decades.

"No other major business type operates under such a high failure rate," writes David Lowe, a medicinal chemist working on preclinical drug discovery. In the end, regardless of whether the truth is closer to 10 percent or 14 percent, as suggested in a survey of clinical success rates across the drug industry done by researchers of the Massachusetts Institute of Technology, it's safe to say that pharma companies that enter the clinic get a pretty low return on their considerable investments.

Drug development productivity — the ratio of the number of new drugs approved to R&D spending each year — has even declined steadily over the past decades. Eroom's Law, as the reverse of Moore's Law, suggests that the cost of developing a new drug has doubled approximately every nine years since the 1950s. Even with inflation rates taken into account, the rise in costs is still tremendous.



Can new technologies help?

Artificial intelligence (AI) is an umbrella term covering a number of technologies, including machine learning (ML) and natural language processing (NLP).

Currently, academic research labs, biotech corporations and technology companies are using machine-based learning to predict pharmaceutical properties of molecular compounds and targets for drug discovery. AI has also proven to be helpful in the enablement of faster diagnosis and tracking of disease progression by using pattern recognition and segmentation techniques on medical images like retinal scans.

However, when it comes to speeding up and improving the success rates for clinical trials, natural language processing offers some very promising opportunities.

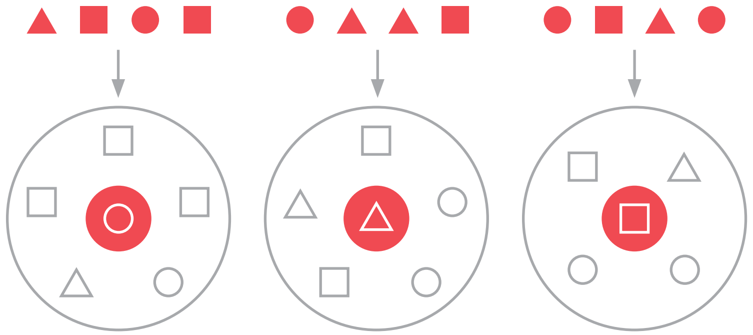
Turning unstructured data into valuable insights

NLP is an area of artificial intelligence and computational linguistics, and essentially a way in which a computer can extract meaning from written or spoken language. It includes information retrieval and extraction, lexical and semantic analysis, pattern recognition, tagging and annotation, and data mining techniques. Rudimentary forms of NLP have existed for many decades. The reason it now offers astonishing benefits is recent progress in statistics, processing speed and the ever-growing amount of available data. Until now, 80 percent of medical data was said to remain unstructured and untapped after it was created.

Modern NLP techniques can help extract valuable information from the ever-growing volume of data. They can be used across virtually any type of textual documents such as electronic health records, clinical trial data, lab reports, white papers, medical and health care regulatory filings, or scientific publications. While a standard keyword search only retrieves documents that researchers must then read, NLP “reads” the documents and can be used, for example, for automated entity recognition, categorization of topic and themes, the summarization of long text bodies or even multi-document summarization, intention detection or sentiment analysis. When extracted in a structured format, these findings enable fast reviews and analysis. For researchers, this often means enormous time savings and an improved basis for decision-making because they can go from finding and reading documents to a data-centric view, uncovering actionable insights from previously hidden relationships.

Relation extraction

Sources like PubMed, ClinicalTrials.gov or Google Scholar contain texts with mentions of authors, clinics, conditions, patients, treatments, adverse events, etc. Using NLP, these bits of information can be grouped together in categories such as people, institutions or drugs to gain a better understanding of their relations.



In drug development, NLP can speed up processes at all stages. Early in target discovery, researchers might search existing specialist literature for recent developments regarding their therapeutic area of interest. Alternatively, they might want to search patient interviews to gain a better understanding of their needs or patent literature to find out more about the competitive whitespace for certain disease targets.

To automate processes, gain deeper insights, reduce risks and lower costs — thereby securing a decisive advantage over their competition — plenty of leading pharma companies have already integrated NLP into their everyday work.

“In the clinical domain, researchers have used NLP systems to identify clinical syndromes and common biomedical concepts from radiology reports, discharge summaries, problem lists, nursing documentation, and medical education documents. Different NLP systems have been developed and utilized to extract events and clinical concepts from text ... Success stories in applying these tools have been reported widely,” states the authors of a research review of clinical information extraction applications that was published in the *Journal of Biomedical Informatics.*

Speeding patient recruitment

A critical area to leverage NLP is the long-standing challenge of poor patient recruitment for clinical trials. Patient recruitment is the largest cost driver of clinical trials. Current estimates suggest that almost 85 percent of clinical trials fail to retain enough patients for successful study conduct. Recruitment and retention-related concerns have been associated with massive delays, with over 90 percent of clinical trials failing to comply with predetermined completion dates, due to poor participant accrual and excessive subject dropout. For a blockbuster drug, such delays can result in capital losses of up to $8 million per day.

The shortage of volunteers may be due to a lack of awareness among the general population, or perhaps issues of trust within the industry. Many patients are either unaware or unsure that participation in a clinical trial is an option at the time of their diagnosis. Moreover, many members of the general public and caregivers lack familiarity with clinical trials and are unaware of opportunities for participation by healthy volunteers. Negative attitudes about participation, which are widely spread, can usually be changed by offering more information.

To be able to contact the right subjects at the right time, clinical research organizations (CROs) need to work closely with physicians, as they are able to offer various treatment options to their patients, including clinical trial research. To find out who is best suited to reach out to potential trial participants, CROs can fall back on a popular approach used when marketing products: the social graph technique. The term refers to a method of data analysis derived from using social networks to find influencers — people engaging with the largest and most relevant audience on social media. It is most often represented as a map with nodes (influencers and followers) connected with lines (various kinds of subscriptions on social media).

To create social graphs useful to CROs, multiple data sources can serve as reference points. For example, they can be developed from or enriched by the data obtained from public datasets, such as PubMed, ClinicalTrials.gov, or H-CUP, and web sources such as Google Scholar, vitals.com, ratemds.com, etc. A great number of medical practices keep anonymized records of their overall patient flow public and many doctors like cosmetologists, nutritionists, and plastic surgeons are even active on social media. With the help of NLP, the data from these datasets can be structured, semantically parsed, and pre-processed with extracted keywords and relationships between nodes.

Ranked by impact and visualized in the form of heatmaps, CROs can easily find the doctors with the most relevant audiences, based on their area of expertise and geographical location, and then advertise to them directly. These doctors can then inform their patients about an opportunity to take part in a clinical trial that could help solve their health issues. As the CROs do not know anything about personal medical details and cannot address the patients directly, the patient’s privacy is preserved at all times.

Identifying top-tier trial sites

NLP can also help improve the chances for successful clinical trials by supporting the evaluation of their feasibility. One of the most important aspects of a clinical trial is selecting high-functioning investigator sites because they can dramatically affect product approval, study costs and timelines. Too often, the identification system for sites is not very mature. As a consequence, the decision of whether a site is deemed suitable is often simply based on the availability of the necessary infrastructure and know-how to fulfill the activities specified in the clinical study protocol. Only one-third of all sites manage to attract enough patients, with many of them falling considerably short.

To improve the process of finding clinical trials that perform well, it makes sense to include criteria such as an investigator’s expert status (e.g., how many articles has he published and how often are they quoted) or prior experience in clinical trials with similar treatments. Other critical factors could be the site’s location and its previous success rates in enlisting subjects, the proximity of comparable studies, or the epidemiological data of the specific patient population. Information like this can be gathered by combining targeted database population, electronic health records, insurance databases, prescriptions, etc. and using NLP techniques to make sense of the data’s semantic relationships.

A semantic relationship could be, for example, asking the solution for sites at which an advanced kind of brain surgery is performed. The system can then gather all relevant sites and a site-scoring algorithm can automatically rank them according to parameters such as the frequency of this special operation, the expert-status of the responsible doctor, the overall site experience with this procedure, or former enrollment rates. The value of this approach is the accurate prediction about the site’s match and the huge time savings for researchers who do not have to do this work manually.

Combining technology and technique

The volume of unstructured data produced in the pharma industry is increasing exponentially, which offers great opportunities. But far too much of the information contained in medicatl records and other documents remains untouched. However, when sorted and analyzed, this data can be used to gain trailblazing insights.

NLP can help to do exactly this and therefore can be helpful during every stage of clinical trials. Combined with techniques like the social graph, for example, it can help meet clinical trial patient guidelines and quickly gather large pools of eligible patients. For this reason, investing in these kinds of AI technologies will not only benefit pharma companies by securing competitive advantages, it will also fundamentally improve the quest for new and better medicines by speeding up clinical trials and reducing their costs.