Unlocking clinical trial data to uncover new therapeutic opportunities

key considerations and best practices

This article discusses opportunities, challenges and best practices in leveraging infrastructure excellence and AI to unlock the value of early or inconclusive clinical trial data. The goal... driving stronger biomarker-based insights and decisions and helping to advance new cures and treatment approaches.

By David Wang, Masha Hoffey and Dr Simone Sharma

As we all know, data is increasingly king in our information-driven world and organisations that learn how to turn mega information into informed strategies are creating true sea changes in their industries. The opportunity to harness the power of data in new and exciting ways is also moving increasingly into life sciences.

In drug discovery, for example, AI can now be employed to mine the mountains of siloed and under-utilised data from early or inconclusive clinical trials. Having this capability helps feed the increasing demand for genotypic and phenotypic data, leads to deeper and faster bio-marker informed insights and helps support the development of new approaches and cures.

Achieving this can seem a herculean task, but there are growing examples of success and best practices around data management, organisational rigour and combining the best of science and technology domain expertise to meet the key challenges posed.

New data insights propel new thinking

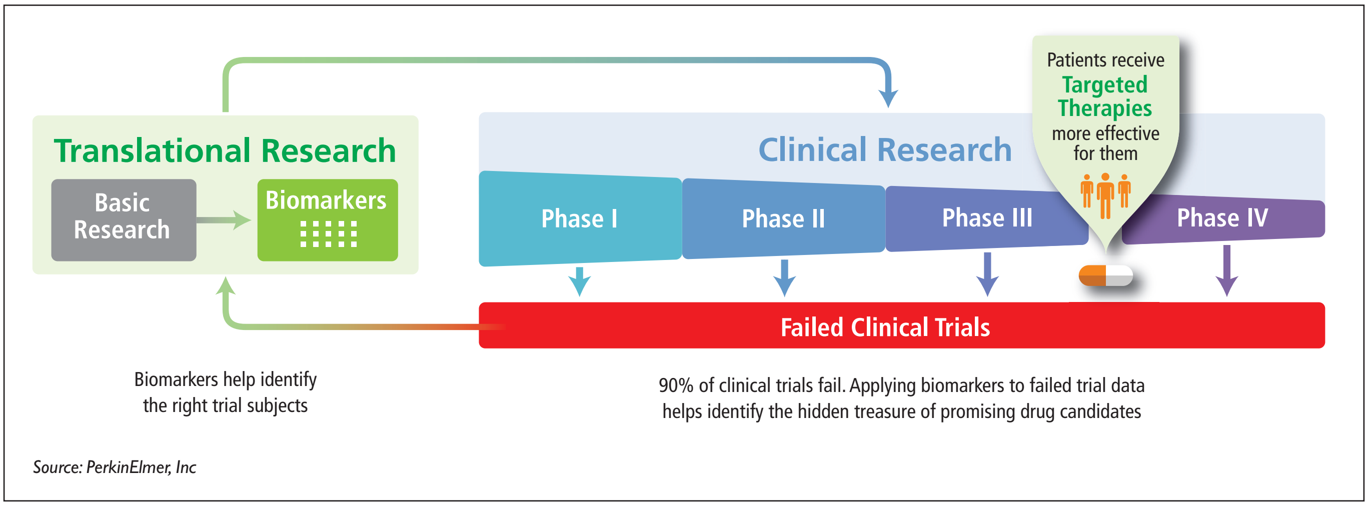
The higher volume and variety of data being generated in our labs certainly brings more data management complexity, but it also brings greater potential for improving the success of clinical trials by better leveraging biomarker-driven science.

For example, routinely-collected biomarker or assay data is now being used to pre-define patient trial subsets around shared, common disease aetiology or molecular profiles. This approach is yielding results and appears to be compelling regulatory authorities, such as the US FDA, to encourage the use of technologies such as advanced analytics for next-generation sequencing and high throughput screening to identify those patients that could benefit most from emerging therapies.

The highly-effective Herceptin® drug, which targets 15-30% of breast cancer patients whose molecular profile indicates higher quantities of the HER2 protein[1], and Merck's Keytruda® drug, are both excellent illustrations of the value of mining clinical trial data to make new advances.

The Keytruda® offering was originally investigated to control the immune response in patients with autoimmune disease. In 2009, Merck's Keytruda® (pembrolizumab) compound was shelved, but promising results from a competitive later-stage compound prompted Merck to take another look at the Keytruda® compound for lung cancer. Stringent selection of patient cohorts using the biomarker-based analytics that resulted, allowed Merck to speed up the drug's approval[2].

The Keytruda® offering is indeed a compelling illustration of the therapeutic value and commercial opportunity that can come from 'decoding' failed


Translational Research
Basic Research
Biomarkers
Clinical Research
Patients receive Targeted Therapies more effective for them
Phase I
Phase II
Phase III
Phase IV
Failed Clinical Trials
Biomarkers help identify the right trial subjects
90% of clinical trials fail. Applying biomarkers to failed trial data helps identify the hidden treasure of promising drug candidates

Source: PerkinElmer, Inc

or inconclusive clinical trial data. The ability to perform retrospective analysis on compound data that has been demoted in priority or placed on an out-license list not only in this case delivered the 'President's Drug', but also led to approval for use in other cancers where treatment is based on a common biomarker versus the anatomical location of a tumour's origin[2,3].

Top challenges in addressing an ailing pipeline with technology

Current technological advances – whether infrastructure-oriented like data storage or around Artificial Intelligence systems – are indeed giving us an unprecedented opportunity to release scientific insights from siloed resources and accelerate the path towards more successful clinical trials.

While there is a consensus in the industry on the benefits of doing this, the challenges can seem daunting. These include:

1) Dealing with high volume, disparate data in corporate data landfills in a scalable, supportable way.

2) Harmonising extracted, relevant data to then perform effective cross-study analysis.

3) Ensuring effective interactive collaboration across teams (within a company and between companies/academic collaborators etc).

4) Staying in lock-step with scientists' needs by identifying what is important in their analysis and then providing ways to intuitively visualise key data.

Even before we think about these issues, however, it is essential to identify and detail the right problem(s) we want to solve before any specific technologies are considered, as biology is complex and insightful analysis can be challenging.

Generally, sponsors appreciate that the key to harnessing therapeutic value in an avalanche of data requires them to be strategic about implementing a system that will address infrastructure and analysis needs so that scientific insights are easily understood and actionable by key decision makers.

Thanks to the pioneering work that has been carried out in other industries by consumer giants such as Amazon and Google, we are now equipped with some solid advances to deal with these challenges in the pharma discovery space.

Best practices are now being established to facilitate data aggregation for cross-study analysis, be it via staging or federated data; creating robust yet flexible models to harmonise datasets; deploying strategies to ensure data is secure by design versus by compliance; or, creating purpose-built workflow solutions that enable analytics for specific end-users without overtaxing highly-specialised data scientists.

Conventional statistical tests work by fitting data into a mathematic model for statistical inference and this can be problematic for complex, high-dimensional data, and especially with historical datasets with many underlying assumptions. Generalised algorithms reliant on a minimal set of assumptions can be much more useful in finding patterns in complex high-dimensional data[4]. Such applications, known as Machine Learning (ML), can help us better leverage and make sense of these data without compromising on the scientific value[5].

Discovering the hidden potential of data with Machine Learning (ML)

High throughput platforms like high content screening (HCS) and next-generation sequencing (NGS) provide data types with millions of features, in which only a small proportion have established clinical significance.

With Artificial Intelligence (AI) systems going mainstream, our ability to find and interpret new information patterns from this complex data and make more reliable and accurate predictions of clinical outcomes has improved drastically[5-7].

Machine learning (ML) is one of the methods to make accurate predictions of future outcomes via pattern recognition to improve patient selection, eg different rates of disease progression. It also helps provide predictive long-term outcomes around safety and efficacy and reduces the time and cost of clinical trials which could vastly improve the drug development process[5,8]. For example, a recent study reported how ML helps detect drug effects that would be missed entirely by conventional statistical tests[9]. Many ML techniques have also been adapted to predict survival or time to progression, but they have not been scalable in performance or interpretation.

Now, using Deep Learning (DL) techniques, cognitive abstraction can be leveraged to identify actionable biomarkers and predict clinical outcomes. The application of such techniques finally enables us to achieve the full potential of siloed data and helps us answer questions that were previously unimaginable.

In addition to helping us make more informed decisions around clinical trial cohort selection, ML can also assist with data preparation such as automating large parts of data curation which is critical to achieving harmonisation across different datasets in trial datasets. Missing historical data can seriously compromise inferences from randomised clinical trials, but ML applications help with imputation of such missing data.

ML can also be used on real world evidence data combined from various sources such as electronic medical records, insurance claims data, prescription data, etc to compare patient populations enrolled in clinical trials for a specific disease. It can also identify potential adverse events that may be correlated to certain unknown subgroup populations[10].

Best practices for leveraging ML

Although the opportunities of ML applications seem limitless, there are also challenges analysing clinical trial data with ML.

While ML success in processing large amounts of structured or image-based data, eg high content screening, is evident, there are many moving parts to clinical trial data, especially segregated or embargoed patient data. These acute challenges increase the complexity for life science companies seeking to leverage ML in Pharma[6,8].

This calls for best practices to be employed. Some include:

Creating cleaner data

Good quality data remains a foundation of strong drug research and development. Since ML applications train on data to improve their predictive modelling capabilities, it is imperative that the data we intend to use is high quality, something which is not always readily available in historical datasets. Messy data (eg inaccurate, incomplete, improperly formatted or duplicated) can significantly distort predictive models. Therefore, organisations need to consider deploying automated tools that not only address data cleansing but also handle data normalisation to make sure that clean and informative data is being fed into ML algorithms.

> "A particular biomarker, for example, can be used to identify appropriate candidates for a clinical trial, such as those patients likely to respond to treatment. This can make it easier and faster to recruit patients and may result in a shorter time for drug approval."
>
> *Janet Woodcock, Director, US FDA Center for Drug Evaluation and Research*

Providing readable/processable data

ML applications are reliant on data that is machine readable/processable. Steps to make data machine readable could include, for example, preparing text data by removing words (tokenisation) which is not trivial when it comes to extracting the right data from medical notes or case report forms (CRFs). Appropriate data parsing, or using curation tools that can automate this process, should also be a primary step in any AI-readiness strategy[11].

Building an appropriate data infrastructure

One of the biggest considerations for running ML applications is the need to have the right infrastructure in place. This involves building an infrastructure that can dynamically scale storage capabilities as data volume grows. It also needs to provide ample computing resources including CPUs and GPUs to support performance and ensure proper data access mechanisms for successful, secure and efficient delivery of data to appropriate users in the organisation. Furthermore, it is important to have high network bandwidth and low-latency work to enable ML applications to operate at an enterprise level[11].

Committing to change management

Enabling meaningful change and ensuring stakeholders adapt to and embrace new technologies can be challenging. Roadblocks range from solutions being rendered too complex or unusable by the intended end-users and convincing end-users in the value of changing traditional data collection techniques[10]. It is vital to clearly and consistently communicate and deliver an aligned direction and strategy, high impact training and clear incentives and benefits. This will help assure that key stakeholders are on board, ready to ask the right questions and committed to making the most of ML's benefits.

Appreciating the limits of ML

Treating any new technology as 'the key' cog within the whole research wheel can cause unforeseen issues. ML applications are no different. Scientific testing relies on a lot of inference where mathematical models are used to test a hypothesis and challenge how the system behaves, and lack of an explicit model can make it difficult to directly relate ML solutions to existing biological knowledge[4]. Although we should leverage our scientific and analytical expertise to move into areas of AI and ML with confidence – we should also be mindful of its true impact and not consider it a panacea for all analytics needs. Treating it as a stand-alone forecasting tool and ignoring all other information from previously established best practices may become counter-productive in the long run. Instead, we should utilise ML applications as an additional and highly-promising aid that will help us better leverage and make sense of the bigger data without compromising on scientific value.

The way forward

Scientific and technological revolutions are well under way and their intersection is poised to give us an unprecedented opportunity to extract more and deeper biological insights from clinical trial data to drive new conclusions, approaches and cures.

We need to harness these revolutions in the most optimal way – leveraging their strong potential but also addressing, head on, the challenges they bring.

The latest estimate for the percentage of drug development programmes that make it from Phase I testing to approval is 13.8% overall and, at 3.4%, oncology drugs have the lowest success rate[12].

Therefore, we need not only to reduce the time to market and cost of new therapies, but we also need to rescue the failing clinical pipeline by embracing innovation.

There has been a huge shift in how clinical trials are beginning to leverage a patient's molecular profile to bring therapeutics into the market that are highly successful for targeted patient populations. Furthermore, the industry is increasingly aware that they must better predict clinical trial success and avoid potential risks. The abilities to meet these demands hinge on how to optimally leverage the historical data sitting in corporate vaults as well as public databases.

As an industry, we need to be able to address the various critical points in the data life cycle, from data collection and access, to data analysis with stakeholder buy-in to truly harness the value of our scientific data investments.

Risks to patient safety and privacy inevitably make Pharma more circumspect when it comes to the application of new technologies. But the industry also recognises the huge potential of these technologies such as ML applications. This promise is realised, as we know, to the point that the FDA is developing a new regulatory framework intended to support the use of ML in clinical trials, drug development and regulatory approval[8].

Furthermore, this is not a one-way street where technologies are informing the data, but the vast amount of disparate and complex data in Pharma is also pushing these technologies to become better and more productive in the long run.

All of this helps ensure that scientific and technological revolutions are not just running in parallel but can truly engage in a symbiotic evolution. DDW

David Wang *is General Manager of Informatics at PerkinElmer. He brings expertise on the transformative value of informatics in delivering smarter decisions and scientific breakthroughs, especially in life sciences. David has held leadership positions at Medtronic/Covidien, J&J and McKinsey. He holds an MBA/Bachelors from the University of Chicago and is completing a Masters in Bioinformatics at Harvard.*

Masha Hoffey *is Director of Clinical & Translational Solutions at PerkinElmer. She leads the development and productisation of solutions that help translational researchers and clinical study teams derive actionable insights from their data. Masha brings more than 10 years of experience in data analytics, regulatory affairs and product portfolio development.*

Dr Simone Sharma *is PerkinElmer's Strategic Lead in Translational Analytics and focuses on driving product direction for data access, integration and advanced analytics of translational research data. She brings a deep expertise in and holds a PhD from University College London in Molecular Genomics.*

References

1 https://web.archive.org/web/20100709180030/http://www.cancer.ucla.edu/Index.aspx?page=36.

2 https://www.forbes.com/sites/davidshaywitz/2017/07/26/the-startling-history-behind-mercks-new-cancer-blockbuster/#23f9e2fe948d.

3 https://labiotech.eu/interviews/interview-keytruda-cancer-inventors/?_sm_au_=iVVJQMNvk44sTJ8FN4s4kKHFLKVG2.

4 Bzdok, D, Altman, N and Krzywinski, M. Statistics versus machine learning. Nature Methods volume15, pages233-234 (2018).

5 Stolzenbach, J. The strive to incorporate machine learning into clinical development. Clinical Trials Arena. https://www.clinicaltrialsarena.com/digital-disruption/the-strive-to-incorporate-machine-learning-into-clinical-development-6218372-2/. Published June 25, 2018. Accessed May 23, 2019.

6 Helfand, C. If pharma looks slow to adopt AI, it's got good reason, expert says. Fierce Pharma. https://www.fiercepharma.com/marketing/if-pharma-looks-slow-to-adopt-ai-there-s-good-reason-expert. Published May 1, 2019. Accessed May 23, 2019.

7 Clinical Research News. Are we ready for AI in clinical trials? https://www.clinicalinformaticsnews.com/2018/12/14/are-we-ready-for-ai-in-clinical-trials.aspx.

8 https://www.clinicalleader.com/doc/machine-learning-in-clinical-trials-what-will-the-future-hold-and-what-s-holding-us-back-0001.

9 https://www.sciencedaily.com/releases/2017/11/171115091819.htm.

10 https://prahs.com/blog/2018/03/23/artificial-intelligence-in-clinical-trials/.

11 https://searchenterpriseai.techtarget.com/feature/Designing-and-building-artificial-intelligence-infrastructure.

12 Estimation of clinical trial success rates and related parameters, Biostatistics 2018, DOI: 10.1093/biostatistics/kxx069.