Data-driven transformation in drug discovery

A review of the current approaches to data management and application within the drug development and research setting, highlighting major critical challenges and emerging solutions for organisations that are determined to harness big data and machine learning.

By Dr Satnam Surae

Data-driven companies that use integrated and advanced analytics outperform their competitors in every sector outside of the pharmaceutical industry. To compete in an increasingly-crowded commercial environment, pharmaceutical and life science companies must gain a greater understanding of the wide-ranging implications of big data and machine learning. These innovations can be applied effectively to drive drug discovery, power research and ensure a sustainable future.

Everyone in the life science sector is familiar with the productivity puzzle: research and development (R&D) spending on drug discovery is increasing, but regulatory approval of new therapeutic agents is largely in decline. Companies are investing more than ever in each new candidate molecules (a 10-fold increase since the mid-1970s), despite widespread awareness that the probability of progressing through clinical trials is less than one in 10 (Figure 1). For some diseases, such as Alzheimer’s, the figure is less than 1 in 100.1,2

Research shows that innovative organisations can optimise the chances of success for their clinical candidates through effective use of these data repositories.3 Efficient data storage and analysis of datasets may accelerate drug development processes. However, existing data management techniques are now struggling to deliver at the scale required to meet the rapidly-increasing quantity of scientific information produced. As a result, pharmaceutical and biotechnology company pipelines are faltering, leaving many businesses unable to effectively manage the mounting pressure on current systems. Specialist platforms, designed to support and continuously evolve alongside drug development and research data outputs, are urgently needed to address these critical industry issues and bring companies into line with environmental demands.

Solutions that harness the rapidly-developing arena of data science will gain a greater competitive advantage. In December 2017, McKinsey described the overall impact of digital technology on R&D as “the $100 billion opportunity”4.

“As we look toward the future of R&D 10 years ahead, we glimpse an entirely new vista: a world where drug discovery is driven by machine learning and advanced analytics mining large data sets, enabling us to understand and visualise interaction with targets and to predict in silico a molecule’s likelihood of success and of reaching approval in the market.”

Meeting the challenge of compliance and complexity
Scientists have turned to ever more sophisticated research technologies to improve their chances of
success in the drug development arena. These technologies have significantly increased the speed and breadth of research processes that may be implemented, generating an exponentially expanding volume of data. Figure 2 shows the growth in genome sequencing data generation alone. However, security and confidentiality issues associated with data access have led to further challenges and complexity, obstructing efficient use of this information and impeding drug development.

There is a growing need for data-driven solutions that support drug development, research and clinical trials, while ensuring that security requirements are appropriately met. Additional regulatory challenges in these fields have also arisen at a time when achieving significant competitive advantage depends heavily on effective data aggregation. For example, collection and aggregation of electronic health records or clinical genomics information, alongside research data, may support identification...
of new drug targets or improve the quality of clinical trial protocols through stratification of patient cohorts. However, use of patient information in this setting (even when anonymised) is subject to strict data protection and ethical regulations that usually require permission to be given on a case-by-case basis. This cumbersome approach precludes routine integration of these data and analyses within R&D activities. These issues are not unique to the pharmaceutical and biotechnology sectors; other industries, such as finance and banking, have also been required to adapt to complex, additional compliance obligations. This represents an important and timely opportunity for companies within the life sciences sector to learn from the experiences of other industries and identify solutions that can be successfully applied within the drug discovery and development environment.

In a typical busy research laboratory, workflows usually span several busy departments. Data from individual experiments are often downloaded to local laptops, or networks, in the form of large Excel spreadsheets. At each stage, the data must be analysed. Each analysis may need to be performed manually over several days and is subject to human error (Figure 3). Complexity is compounded by variation in the way that data are structured (or even named) across a workflow. Breakthroughs can occur when trends, patterns and relationships are identified within the data. Unfortunately, inconsistencies and inefficiencies within existing systems mean that data, which should be an asset, becomes an impenetrable liability. Given the size of many research organisations and the scale of work undertaken within each department, it is easy to see how productivity and quality may be improved by data management systems that offer consistency and a defined structure.

Creation of one common platform that aggregates, structures and digitalises workflows across an organisation can unlock the potential of data, providing easy access and opportunities for greater and more productive cross-department collaborations. Once data is properly structured, analyses can be customised and performed at the touch of a button.

**Embracing a data-based revolution**

Specialist expertise that has traditionally only been available to universities and other academic institutions, is now emerging through a new generation of start-up companies. Companies that harness this academic knowledge are poised to make breakthroughs, not just in pipelines and processes, but in the fundamental way that businesses control, shape and steer data flow through their entire organisation. These changes are so radical and happening so fast that companies will need to adopt an entirely new perspective concerning their understanding of the data they produce and the way that it is applied within the research or commercial environment.

R&D-focused organisations within the pharmaceutical and biotechnology industries are placing greater emphasis on becoming more literate with big data technologies. Large corporations, such as Novartis, are reinventing themselves as ‘medicines and data science’ companies. This has raised the profile and importance of bioinformaticians within organisations who have traditionally provided service roles for biologist colleagues concerning data analysis, statistics and pipelining of tools. The value of bioinformaticians, especially those with an understanding of the challenges of commercial R&D, is increasing in line with the rapidly-evolving demands of the environment. Bioinformatics is
In informatics, fast becoming a core capability for these organisations across the R&D disciplines, from discovery, validation, pre-clinical and translation trials, through to clinical trial design. The industry now recognises bioinformatics data science expertise as a competitive advantage that will allow businesses to gain maximum value from both their in-house and public data sources.

This paradigm shift is long overdue. The revolution is based on carefully building a closer interaction between the scientists that make research decisions and the people who build their technological solutions. Bridging these fields brings the promise of change and leading-edge innovation throughout the sector over the coming years.

**What does it mean to be truly data-driven?**

We are only just beginning to understand and appreciate the benefits that data management can provide for the pharmaceutical and biotechnology industries. The sector is on the verge of discovering new concepts that will allow organisations to steer and surf the flow of an emerging new type of dataset: a vast, all-consuming, constantly changing dataset that permeates deeply through an organisation and beyond.

Data management and machine learning, when it is ‘baked’ into a data science platform from the outset and built from the ground up by life science experts, has the most realistic chance of rising to this challenge. This approach is not constrained by previous logical, structural (and even linguistic-emotional) constraints. Finding new ways to understand and address data management challenges, in partnership with the researchers who are working in a high-pressure drug development environment, can help to conquer these restrictions.

Within life science organisations, data is regularly siloed and unable to flow through departments and groups. Repetitive analytical tasks lead to frustration and absorb time because they are labour-intensive, instead of being executed automatically. Traditionally, analytical pipelines for drug discovery are developed and executed using data to answer biologically pertinent questions. This is accomplished laboriously by hand, requiring huge resource investment and constant validity checks to ensure quality. There is a lack of feedback and agility, and an inability to forecast progress. As a result, organisations can miss out on finding the drug candidates that present the best chance of commercial success. Most of these problems can be traced back to data analysis systems that are simply not fit for purpose at the current time and present a major barrier in terms of long-term pipeline success.

In a cutting-edge life sciences environment, it is essential to create effective structures that align with the demands and challenges of research and development. Data must be structured effectively if machine learning and advanced analytics are to play a meaningful role in the drug discovery process.

Diverse data sources and results from multiple in-house experiments must be integrated and enriched with large external data repositories. For example, robust validation of hypotheses underpinning important R&D pipeline work...
may be enabled through integration of specific in-house generated research data with large and more diverse public data repositories from initiatives such as UK Biobank or the 100,000 genome projects.

**Machine learning: driving informed decision making**

The process of addressing universal data management issues and tackling common environmental difficulties has led to the successful automation and streamlining of diverse analytical processes. New machine-learning technologies allow datasets to be brought into a truly data-driven decision-making process. These datasets may span a number of varied workstreams, including mass spectrometry, next-generation sequencing (NGS), high-throughput imaging, immunoassays and biophysical assays (Figure 4). This helps to empower both the scientists and the companies they work for to ensure that those candidates that are brought forward to clinical trials are supported by a comprehensive and robust pre-clinical data package and have the greatest probability of success.

With the right environment, decision making becomes faster and of higher quality. Scientists have greater clarity concerning the scientific evidence available to them, without needing to manually amalgamate disparate data sources. For example, emergence of adverse drug reactions or side-effects may be a key reason that clinical trials fail during candidate selection. Integration of diverse data assets, in combination with a variety of predictive tools, increases the likelihood of finding these side-effects at an earlier stage in development, before a new drug researches the clinical phase.

McKinsey highlights this in its vision of the data-driven digital future for biotech.

“Insights from in silico studies and analysis of diverse datasets [will] accelerate research and early development through more informed decision making, including smoothing the repurposing of existing drugs for new therapeutic areas.”

Platforms with a strong machine learning element can identify patterns and trends, beyond those visible to the human eye. Highlighting these patterns will assist scientists in focusing on the important data, cutting down on distracting ‘noise’ or irrelevant information. Time and effort are decreased because the number of experiments required is reduced. There may also be situations when information from candidates that have previously failed may inform decisions regarding the development of new candidates with similar data profiles. This allows some candidates to ‘fail fast’ as development programmes may be cancelled at an earlier stage to avoid needless investment and analysis. The fail fast paradigm is economically critical to drug developers because higher costs are generally incurred as candidates move closer to clinical trials.

**Digital transformation in data management**

From the bench-top to the boardroom, there are unique challenges that come as part of the inevitable digital transformation in life sciences and drug discovery. Any successful response to these challenges must be modular and continuously adaptable to the unique and particular research processes and structures of an individual organisation as it grows.

New platforms are now meeting the needs of research-intensive organisations in the life sciences sector, leveraging cutting-edge IT concepts such as the cloud, DevOps and software-as-a-service. The most successful are employing machine learning and artificial intelligence (AI) technologies. These platforms are already making an impact, with immediate benefits for researchers, IT departments and management departments; improving productivity exponentially.

Security is another pivotal consideration for data management platforms and should not be implemented as an after-thought. Standard security certifications such as ITIL, ISO:9001 and ISO:27001 should be sought. Compliance with GDPR is also critical when dealing with personally identifiable data, whereas relevant GxP/FDA/EMA regulations for data analysis or processing are only needed when the output of these data platforms contributes directly to manufacturing processes, companion diagnostics or clinical trial design/applications.

**Future perspectives and powerful partnerships**

Innovative data management platforms enable huge and diverse datasets to be structured, stored and analysed effectively. Drug developers who fully embrace machine learning and AI technologies will make rapid, impactful and significant leaps in their research. Leveraging these platforms will lead to the emergence of novel data patterns and ideas that inspire new approaches to data interrogation and research. This will, in turn, reveal fresh scientific challenges that may advance the fields of medicine and biotechnology in unexpected and exciting ways that would not have been previously possible.

Not all pharmaceutical and biotechnology
companies can boldly reinvent themselves in the same way as Novartis. Fortunately, the majority of organisations may be able to reap the rewards of effective data management and digital transformation without needing to make significant structural changes. Working in close partnerships with specialist technology providers can enable companies to benefit from their technical expertise and apply this knowledge across the research environment. This allows pharmaceutical and biotechnology companies to fully focus on the output of their drug discovery and development pipeline, directing their resources appropriately and cost effectively to bring better drugs to more patients in less time. With the right support, companies can sharpen and refine their commercial strategies to make the best use of the wealth of data that they have at their fingertips, ensuring that they are fit for the challenges of today and the future.

Any dynamic and innovative life sciences organisation must now become part of this data revolution. At Aigenpulse, we believe the winners in this total digital transformation will be those who adopt agile approaches, empower scientists and enable data-driven technology throughout their entire research and development process. DDW

References


Dr Satnam Surae has been active in the life sciences sector for more than 10 years. Satnam received a BSc (Hons) in Biochemistry and MRes in Computational Biology from the University of York (UK), and a PhD in Computational Structural Biology from University College Dublin (Ireland). Subsequently, he worked in industrial biotechnology developing metabolic models, prior to joining Aigenpulse.