Life Sciences Companies Must First Solve Data Silos to Enable AI's Potential

May 23, 2024

HCC was founded in 2015 to invest in companies applying AI and other information technology in healthcare and related insurance and financial services adjacencies in ways that improve the human condition and deliver better value. Recent advancements in AI increased our interest level in finding new applications for AI in healthcare. These advances include improved natural language understanding, more accurate and context-aware responses, and the ability to perform a wider variety of tasks such as writing, translation, and answering questions. These advances are made possible largely due to increased computational power and refined training techniques.

Coming back to shore from an extensive deep dive into the applications of information technology and especially AI in the pharmaceutical industry, we share with our investors and others some high-level knowledge and near-term conclusions we gained from our primary research.

The life sciences industry is of particular interest because of the size and strategic relevance. And to HCC, it's especially intriguing as a well-funded frontier for applied AI.

Over the past several decades, the life sciences industry has faced a troubling decline in R&D productivity. The cost to bring a new drug to market has increased from about $400 million in the 1980s to over $2 billion in the 2010s, adjusted for inflation.[1] This cost increase is largely due to increasingly complex biology and the inefficiency of the traditional drug development process that relies on a "trial and error" approach. AI stands out as a potentially transformative technology for life sciences companies, capable of analyzing vast amounts of complex data and predicting experimental and clinical outcomes. NVIDIA CEO Jensen Huang has even stated that digital biology will be one of the most significant revolutions in human history.

And yet, the life sciences industry is plagued by a problem preventing it from unlocking the full potential of AI, a problem that other parts of the healthcare ecosystem know all too well: siloed data. According to a McKinsey survey, lack of integrated data sources was cited by pharma leaders as the largest hurdle to scaling digital and analytics capabilities.[2] Furthermore, nearly half (48%) of senior decision-makers in pharmaceutical companies surveyed by Aspen Technology stated that data silos derailed the efficiency of cross-functional collaboration in their organization.[3]

Data is often siloed within different departments, systems, and repositories in life sciences companies. Without an integrated infrastructure, it's difficult to pinpoint existing data, access it quickly, and leverage it across the organization, which often stifles decision-making and innovation. Data is the cornerstone of AI models and remains a top priority in the life sciences AI race. A key strategic differentiator is the ability to generate, acquire, and control large datasets from diverse sources for training models in areas such as omics, imaging, and health records.[4]

Before life sciences companies can achieve digital and analytics transformation, they must embrace data transformation, which is where DataOps comes in.

- DataOps is a set of collaborative practices, capabilities, and tools that can standardize and automate data use to improve quality and reduce the cycle time of advanced analytics.[5]
- Its framework includes data integration, data management, analytics development, and data delivery.

- DataOps offers a myriad of benefits to life science organizations, including enhanced data quality, boosted efficiency, strengthened collaboration, robust data governance, and accelerated time to insights.[6]

From our primary and secondary research into the life sciences DataOps landscape, we've noted six key trends:

1. Life science companies aspire to integrate all generated data, but the industry is still early in its journey. Efforts currently focus on different steps of the drug value chain, tailored for each end user, resulting in fragmented solutions for R&D, clinical trials, commercial operations, and manufacturing.

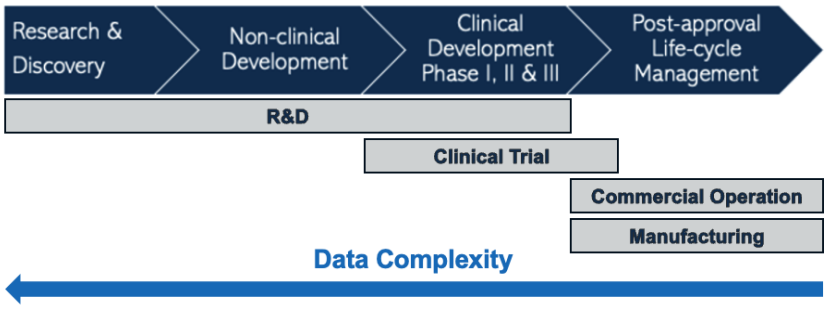

Exhibit 1: Segments by end user in the drug development process.

2. Data complexity is particularly pronounced at the early stages of the drug value chain, especially in early drug discovery. This complexity has spurred greater innovation and the rise of hundreds of startups aimed at solving data silo issues.

3. The past two years have been challenging for the life sciences industry with both business obstacles - rising interest rates, layoffs, and a slow fundraising environment - and scientific obstacles - increasing complexity of biological systems, limited identification and validation of new drug targets, and drug resistance. These challenges have made it difficult for small biotechs to invest in data services while cutting operating budgets. In contrast, large pharmaceutical companies continue to invest heavily in R&D, totaling over $145 billion for the top 20 pharmaceutical companies in 2023.[7]

4. Both small biotech firms and large pharmaceutical companies are allocating more resources to later-phase drugs, indicating a strong demand for data services in late drug development, clinical trials, commercial operations, and manufacturing. As a result, DataOps startups focusing on early drug discovery face adoption challenges despite the long-term promise and importance for the industry.

5. The life sciences industry has vast amounts of unstructured data. The development of large language models (LLMs) presents new opportunities to analyze and integrate this unstructured data with structured data. Consequently, there is a growing demand within the sector for tools that can efficiently process unstructured data.

6. A DataOps company can set itself apart from the competition by offering more than just tools; it can provide proprietary data as well. Possessing exclusive datasets enables clients to uncover insights that aren't accessible through public or widely used sources, presenting a significant opportunity for upselling.

So, what are the characteristics of a best-in-class DataOps platform for the life sciences industry?

- The ability to collect and aggregate data from a wide range of sources.
- The ability to integrate external data with internal data.
- Ease of integration with existing products and platforms.
- The use of powerful large language models (LLMs) trained on an organization's own data.
- Robust cybersecurity features to ensure data safety and privacy.

We remain optimistic that the life sciences industry can harness the power of AI to accelerate its development of new medicines, but it must first develop more internally integrated data systems to unlock the full potential of AI.

In the course of our primary research, we discovered over 500 companies that claim to be using AI in the drug discovery process. We continue to search for promising businesses to partner with and support in a scale-up.

Sources:

1. Nature Reviews Drug Discovery, *Diagnosing the decline in pharmaceutical R&D efficiency*, March 1, 2012.
2. McKinsey, *Top ten observations from 2022 in life sciences digital and analytics*. January 31, 2023.
3. European Pharmaceutical Manufacturer, *Data silos threaten efficiency levels for nearly half of pharma businesses*, February 13, 2023.
4. BioPharma Trend, *It's been a decade of AI in the drug discovery race. What's next?* April 3, 2024.
5. McKinsey, *Rewired pharma companies will win in the digital age.* June 14, 2023.
6. Airbyte, *DataOps: the definitive guide to streaming data pipelines.* May 5, 2023.
7. Deloitte, *Measuring the return from pharmaceutical innovation – 14th edition*. April 2024.