tetrascience

White paper

Challenges and Future Directions in Chemistry, Manufacturing, and Controls (CMC)

Ken Fountain[1] and Mark Buswell[2]

[1]TetraScience, Inc., Boston, MA; [2]NGT Biopharma Consultants, London, UK

Table of Contents

Abstract

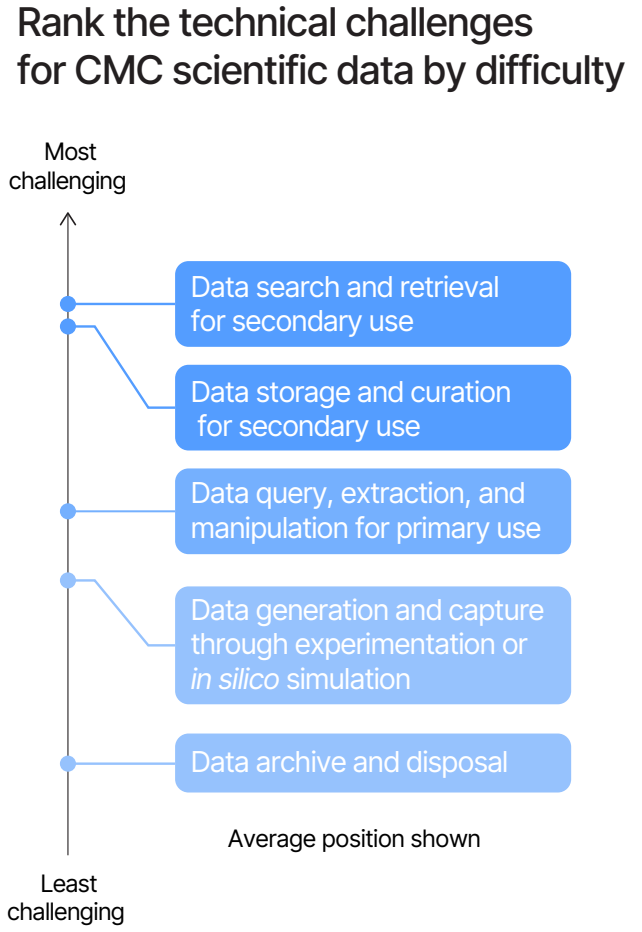
This white paper summarizes the key findings from a group of industry experts on the challenges and future directions of scientific data in CMC. The top key business challenges facing the industry are late-phase portfolio acceleration, CMC resource efficiency, and a shift toward electronic regulatory submissions starting as early as 2026. Data lifecycle challenges associated with these business challenges are almost exclusively related to storage, curation, search, and retrieval of CMC data for secondary use. The biggest and perhaps most complex challenge comes from the people and culture aspect; lab scientists are still entrenched in a paradigm based on traditional paper notebooks for recording experiments. They primarily focus on the immediate experimental task and outcome, rather than considering the downstream usage of their data. The reasons for these challenges, along with the industry's general readiness for upcoming regulatory changes for electronic drug filing submissions, are explored in detail. Opportunities for leveraging advanced analytics, artificial intelligence (AI), and machine learning (ML) to solve the key challenges with scientific CMC data and future regulatory filings will be covered.

Introduction

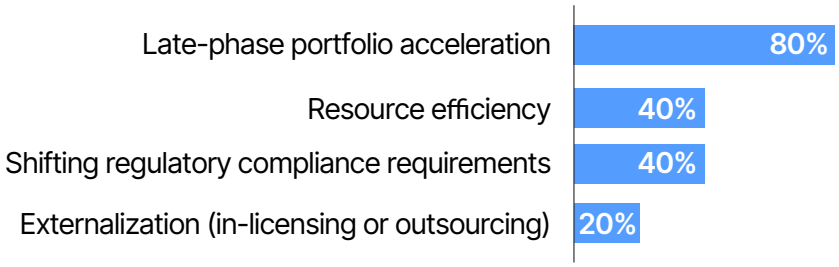
The field of chemistry, manufacturing, and controls (CMC) in pharma is integral to the drug discovery and development process, focusing on the design and production of pharmaceutical products. This encompasses delivering the active ingredient, ensuring the reliability, safety, and scalability of manufacturing processes, and maintaining the quality and consistency of the drug substance and product to guarantee patient safety and efficacy. In essence, pharma CMC functions generate two outputs: clinical trial materials for the various phases of clinical trials and CMC data and knowledge. The latter supports regulatory submissions to achieve approval for a new therapy and facilitates technology transfer of the manufacturing process. As such, the generation, management, and use of CMC scientific data is a critical topic for CMC functions. This white paper discusses the challenges and future directions related to CMC scientific data based on the input of a CMC subject matter expert (SME) group.

CMC business and data lifecycle challenges

There has been significant technology investment in drug discovery over the past five to ten years, as the industry has sought to exploit genetics and functional genomics to improve the quantity and quality of targets and early-phase assets.[1] The success of this strategy is now putting pressure on CMC functions to accelerate these promising pipelines through to file and launch with shorter timelines and reduced costs. Our expert panel identified the top three business challenges facing CMC functions as late-phase portfolio acceleration, resource inefficiency, and a shift toward structured data for electronic regulatory submissions.



What are your top two business challenges in CMC over the next three years?



While late-phase portfolio acceleration was widely seen as the most important business challenge, the nature of this challenge depends on modality and whether assets are internally discovered and developed or in-licensed from early-phase biotechs. Traditionally, for large pharmas developing small molecules, CMC functions have prided themselves on not being on the critical path for late-phase acceleration. This was achieved by addressing CMC early with adequate resourcing. The rise in late-phase in-licensing of assets with underdeveloped CMC packages, coupled with the push for greater resource efficiency, places pressure on small-molecule CMC functions to stay off the critical path while still maintaining the quality of CMC outcomes. These challenges are heightened with large-molecule biologics, vaccines, and cell therapies due to the relatively complex nature of the drug substance itself, as well as the biological processes used to produce the drug substance. As such, CMC functions are often under even greater pressure with biologically derived modalities.

A major obstacle in using technology to overcome these business challenges is the persistence of traditional, document-centric CMC workflows. They are characterized by manual transcription and duplication of data across numerous technical documents, which leads to inefficient decision making and suboptimal CMC development plans. Documents are seen as the source of truth, rather than source systems or data platforms. The lack of automated data flow exacerbates this problem, as critical information remains trapped within primary systems, or worse, within documents, rather than seamlessly flowing through line functions. Without a concerted effort to modernize CMC workflows and processes and move away from the manual document-centric paradigm, the industry will continue to experience a data drag that erodes efficiency, prolongs timelines, and diminishes the quality of CMC outcomes.

Factors contributing to CMC data drag include the lack of an industry-standard data model for CMC and the absence of equipment connectivity and data curation. These issues make it hard to streamline operations for efficiency and speed and to enable data to flow. The lack of connectivity between scientific equipment and lab informatics systems, in particular, hampers efforts to achieve better efficiency through digital labs and generates challenges for data integrity and real-time monitoring. Without a driving force to address these inefficiencies, CMC functions will struggle to keep pace with evolving regulatory and operational demands, impacting their organizations' overall productivity and competitiveness.

These challenges are even more acute in the context of late-phase asset in-licensing. Typically, in-licensed assets are often underdeveloped from a CMC perspective, introducing considerable risk into CMC operations. The lack of data standards is a challenge within and, even more so, between companies. When in-licensing an asset or outsourcing work to contract organizations (e.g., CROs, CDMOs, and CMOs), it is difficult to integrate R&D or CMC data into the receiving company's system. With the reliance on late-phase in-licensing likely to increase, CMC functions need better tools and systems to ingest data and quickly identify risks.

In the data lifecycle, data storage and curation, as well as subsequent data search and retrieval, pose the most significant technical challenges for CMC SMEs. There are several contributing factors. First, processes within CMC are often designed to generate data for immediate use rather than secondary purposes. This approach creates a disconnect between data generated at different stages of development, hindering the ability to bridge critical information across various areas of CMC. One example cited was the inability to effectively link early predictive stability data with actual stability results later in the process. This impacts the identification of key drug substance (DS) or drug product (DP) characteristics early in the development process.

> **Processes within CMC are often designed to generate data for immediate use rather than secondary purposes.**

Second, managing CMC data is significantly complicated by the diverse lab informatics landscape and the lack of a standardized data model for formatting and storing CMC content. While efforts like FDA PQ-CMC, ISO Identification of Medicinal Products (IDMP), and the upcoming International Conference on Harmonization (ICH) Structured Product Quality Submissions (SPQS) aim to address this issue, the adoption of a common CMC data model remains incomplete. To bridge this gap, companies often rely on custom point-to-point solutions and subject matter expertise for data aggregation.

Third, people and cultural aspects contribute to the challenge. Despite the deployment of electronic lab notebooks (ELNs) and laboratory information management systems (LIMSs) in pharma CMC development labs, many scientists still record experiments in a traditional paper notebook and then generate static reports to aggregate and summarize their experimental work. This paradigm is partly due to the first generation of ELNs being conceived as digital counterparts to paper notebooks instead of tools to enable a fully end-to-end digital workflow for achieving CMC outcomes. Understandably, many lab scientists primarily focus on the immediate experimental task and outcome rather than the downstream usage of their data. Attempts to shift their focus often impose manual data entry requirements when setting up experiments. This burdensome process significantly erodes the user experience and slows the adoption of digital workflows and systems that would otherwise address data lifecycle challenges.

There is a pressing need for a new generation of workflows, systems, and system architectures designed by first intent for both primary and secondary CMC objectives. Such advancements would automate data assembly and contextualization and facilitate the end-to-end flow of CMC data. They would also ensure the availability and accessibility of CMC data and knowledge for effective communication with regulators and seamless integration into manufacturing operations.

Preparedness for electronic regulatory submissions

While the challenges discussed so far are internally focused, there are shifting external regulatory expectations that will demand a response from CMC functions. Specifically, regulatory agencies are beginning to introduce fully digital electronic submission requirements for certain CMC data sections. All participants of our expert panel were fully aware of impending electronic submission regulations, but some companies have yet to fully prepare. Initiatives like ISO IDMP have paved the way for the full structuring of the entire CMC section, indicating a positive trajectory toward electronic submissions. However, the absence of an international data standard impedes progress. As a first step, pharmaceutical companies are actively engaged in working groups with regulatory bodies to define an international data model. A significant hurdle still remains: the need for a shift in mindset toward a standardized data approach that extends all the way back to the source systems generating the data.

The industry envisions a streamlined experience in the future, where the automated flow of CMC data facilitates the automated authoring of structured content for regulatory data submission. This will require tools and systems for selecting data, employing automated table builders, and seamlessly managing and integrating CMC content directly with regulatory submission portals. Such an approach promises to eliminate data integrity checks and promote the reuse of components across various documents, enhancing efficiency and reducing redundancies. From an industry perspective, achieving this shift in a synchronized way across multiple regulatory authorities will be key. Despite the benefits of this trend, it was acknowledged that validating these automated systems and processes for structured content authoring poses a significant challenge. This is especially true for smaller players (i.e., smaller biotechs), where the costs associated with implementation may prove prohibitive due to limited filing frequency.

Governance challenges in CMC scientific data

Despite the potential benefits of data technology, CMC functions face several challenges in securing investment for scientific data solutions. First, it is often difficult for CMC SMEs to explain the return on investment to senior management. The costs of implementing such solutions are often prohibitive when compared with the perceived value of investment in terms of portfolio acceleration and productivity enhancements. Drawing a thread between a specific technology investment and tangible outcomes can be tricky, especially when multiple initiatives and programs are implemented concurrently to address acceleration and productivity. Yet another challenge lies in crafting a compelling business case. There is a strong bias toward investment in tangible short-term returns, making it difficult to justify investment in long-term, more strategic capabilities. Advocates for CMC data technology investment and vendors of platforms and solutions will have to partner much more transparently and purposefully to address these concerns and secure investment.

Second, and regrettably, previous investments in CMC data technology have often resulted in underestimation of implementation and lifecycle costs, lack of tangible return after implementation, and considerable challenges with system adoption. Such experiences foster a natural skepticism and caution toward future technology investments.

Opportunities for AI/ML in CMC

The opportunities for AI/ML in CMC present a promising avenue for revolutionizing the pharmaceutical industry. The key to realizing this potential lies in the availability of high-quality training data, contextualized to specific CMC processes. This data serves as the foundation for leveraging AI/ML algorithms to model scientific processes with precision and efficiency. By modeling processes by first intent, organizations can minimize the need for costly and time-consuming laboratory work, particularly in the development of biologics and cell and gene therapies (CGT), which are inherently more complex than traditional small-molecule modalities. However, such data is often fragmented across different organization groups, requiring substantial resources to compile. Industry-wide collaborative efforts to share CMC data could significantly enhance the quality and quantity of training data available, unlocking the full potential of AI/ML applications in CMC.

> **By modeling processes by first intent, organizations can minimize the need for costly and time-consuming laboratory work.**

The industry must redefine the single source of truth for scientific data to fully capitalize on AI/ML capabilities in CMC. As already discussed, many organizations rely on PDF documents as the definitive source of truth, which are manually corrected and curated by SMEs. However, to realize the transformative potential of AI/ML, the focus must shift toward structuring data and making it readily accessible within source systems. This necessitates robust data lineage, standardization, and quality control processes to maintain data integrity throughout its lifecycle. By establishing structured and easily accessible data repositories, organizations can streamline processes, enhance decision making, and better utilize AI/ML technologies.

The first generation of LIMS, ELNs, and chromatography data systems (CDSs) played a big role in digitizing labs. Nonetheless, these systems must now evolve to facilitate fully digital workflows. They need to promote a data-centric mindset and improve the user experience of capturing and creating high- quality structured and contextualized CMC data. Many existing data tools excel at handling structured data but are underutilized due to cumbersome user interfaces and complex workflows. Enhancing the usability and accessibility of these tools can empower scientists to efficiently generate high-quality structured data, facilitating the seamless integration of AI/ML technologies into CMC processes.

Moreover, leveraging automation to create secondary-ready structured data can further enhance efficiency and accuracy, reducing the burden on scientists and enabling them to focus on higher-value tasks. By addressing these technological challenges and implementing the necessary change management, organizations can unlock the full potential of AI/ML in CMC and drive transformative change within the pharmaceutical industry. Examples include using machine learning to optimize complex biological processes to enhance quality and yield, leveraging digital twins in process control to optimize manufacturing quality aspects, and using generative AI to generate protocols and SOPs. No doubt there will be many other use cases as CMC departments further explore these technologies.

Conclusion

While CMC is critical to the overall drug discovery and development process, many CMC functions are struggling to break out of a document-centric data-silo paradigm. This dilemma persists despite shifting regulatory expectations and clear opportunities to enhance CMC outcomes by leveraging advanced technologies. Our expert group identified four key future directions: (1) enhanced end-to-end flow of CMC data, (2) automated synthesis of CMC data linked to in-licensing, (3) automated contextualization of CMC data during generation to minimize the burden on scientists, and (4) creation of a standardized CMC data model. However, to pursue these directions, CMC SMEs and vendors must work together to explain the opportunities more effectively and make clearer commitments on the value and return of investment. Pharma organizations will need to embrace a data-centric mindset via change management and deploy technology to lab scientists to improve data quality and reduce the burden of capturing context. Despite these challenges, investments in CMC scientific data management offer clear opportunities to accelerate CMC, improve efficiency, and ultimately accelerate medicines to patients.

Acknowledgements

The authors would like to thank the following industry experts for their contributions to this white paper: Timin Hadi (Amgen), Oliver Hesse (Bayer Pharmaceuticals), Femi Akintobi and Matt Harrison (GlaxoSmithKline), Gary Woo (Landmark Bio), Laurent Lefebvre (Novartis), Yuanqi Tao (Takeda Pharmaceuticals).

References

[1] M. Schuh, F. J. (2019). Advances in genomics for drug discovery. Drug Discovery Today, 24(2), 534-545.