Paper PD07

From Data Manager to Data Scientist: Navigating the Future of Data Management in Clinical Trials

Debra Anderson, Veramed, Boston, USA

ABSTRACT

The role of Clinical Data Management (CDM) is changing to meet new challenges such as decentralization of clinical trials, new data sources, real-world data, and evolving data management environments. Clinical Data Science (CDS) represents the adoption of new technologies, new processes and new data analytics skills while also continuing to maintain a deep understanding of the clinical trial domain, good clinical data practices, and other core CDM competencies which are essential to support efficient decision-making throughout the life cycle of clinical research. This paper explores the professional development challenge in transforming from CDM to CDS in terms of foundational knowledge, mindset, technical, and soft skills.

INTRODUCTION

The clinical trial landscape is evolving rapidly, driven by advancements in technology, the growing variety of data sources, and the increasing complexity of studies. The role of the Clinical Data Manager, once primarily focused on managing and cleaning clinical trial data, is now shifting toward the more dynamic and analytical area of Data Science. In this paper, I'll explore how the Clinical Data Manager role has changed, the impact of new clinical data sources, the integration of emerging technologies, and strategies to prepare for professional success in this new environment.

CHANGING ENVIRONMENT OF CLINICAL TRIALS

Clinical Data Managers have been responsible for tasks such as designing case report forms (CRFs), development of data collection systems, maintaining data quality, coordinating external data vendors, and ensuring compliance with regulatory standards. These responsibilities require strong organizational skills and a deep understanding of clinical trials. While these core duties remain essential, the role of the Clinical Data Manager has expanded significantly in recent years.

With clinical trials becoming more complex, Clinical Data Managers are now expected to have advanced analytical skills and contribute to data driven decision making. The rise of electronic data capture (EDC) systems and standards such as CDISC SDTM and ADaM added a more technical dimension to the role. Clinical Data Managers are starting to work more closely with biostatisticians, programmers, and regulatory teams to ensure that data aligns with broader trial objectives.

As industry shifts toward decentralization and real-time data collection, Clinical Data Managers are stepping into Clinical Data Science. According to the SCDM's position paper, this evolution includes embracing foundational data science principles such as data wrangling, machine learning techniques, and advanced data visualization to gain actionable insights from diverse datasets [1]. It's an exciting time that offers Clinical Data Managers the chance to expand their skillsets and make a greater impact on clinical research.

Data review is an essential part of monitoring clinical trials. It goes beyond on-site monitoring and involves overseeing the entire trial. ICH E6 guidelines emphasize that sponsors must develop a risk-based monitoring approach, which could include on-site, centralized, or a combination of both methods [2]. Clinical data review fits here, focusing on timely remote evaluations by trained experts like Clinical Data Managers.

Regulators have shifted from demanding perfect data across the board to prioritizing critical data that is accurate and fit for purpose. Sponsors are aligning with this shift by updating their data review strategies. This involves not just patient data but also audit trails, which reveal how the study is conducted and data collected. Reviewing these trails helps ensure data integrity and quality, especially for third-party sources like eSource data. Success in this area requires the right tools, including intelligent technologies, to enhance data reviews.

EXPANSION OF CLINICAL DATA SOURCES

The sources of clinical trial data have grown dramatically. In the past, data primarily came from CRFs and laboratory results. Today, we are seeing data from wearable devices, electronic health records (EHRs), real-world data (RWD), and patient-reported outcomes (PROs). While these new sources bring incredible opportunities, they also introduce challenges in terms of data integration and quality.

- **Wearable Devices and Sensors**: Wearables provide continuous streams of data, offering unprecedented detail in patient monitoring. However, these devices generate vast amounts of unstructured data that require advanced tools to process.
- **Real-World Data (RWD)**: Data from sources like insurance claims, registries, and prescriptions add valuable context to clinical trial outcomes. Integrating this data requires robust harmonization and standardization techniques.
- **-Omics Data**: Genomic, proteomic, and metabolomic data are becoming more common in clinical trials. Managing these datasets demands specialized knowledge and computational resources.

As these data sources expand, ensuring interoperability and data governance has become more critical than ever. Clinical Data Managers must now ensure diverse data sets can be seamlessly integrated while maintaining data quality and regulatory compliance.

TECHNOLOGY AND ITS ROLE IN SHAPING THE FUTURE

Advancements in technology are revolutionizing Clinical Data Management. Several innovations are changing how we collect, analyze, and report data in clinical trials:

- **Artificial Intelligence (AI) and Machine Learning (ML)**: These technologies enable predictive analytics, anomaly detection, and automated data cleaning. AI can identify patterns in data that predict patient outcomes or highlight protocol issues. Data cleaning activities previously completed by Clinical Data Managers are now able to be completed by ML algorithms.

- **Cloud-Based Data Platforms**: Cloud solutions allow real-time data sharing and collaboration across teams in different locations, improving scalability and speeding up decision-making.

- **Visualization Tools**: Tools like Spotfire, Tableau and Power BI make it easier to interpret complex datasets. Interactive dashboards help teams monitor trial progress and make informed decisions.

Leveraging these technologies requires a mindset and skill shift. As these technologies become more widespread, Clinical Data Managers must develop technical skills to use them effectively.

EXPANSION OF DATA SCIENCE UNDERGRADUATE PROGRAMS

Over the past decade, data science programs have seen remarkable growth in higher education, reflecting the increasing demand for professionals skilled in data analysis and interpretation. In 2020, the National Center for Education Statistics introduced new Classification of Instructional Program (CIP) codes specific to data science and analytics, facilitating more accurate tracking of these programs [3]. Between 2012 and 2021, data science and analytics programs experienced a staggering 700% increase in completions, rising from just under 6,000 to over 46,000 [3]. This surge is particularly notable at the master's level, which accounts for 68% of these completions, indicating a strong emphasis on advanced education in this field [3].

The rapid expansion of data science programs is also evident at institutions like the University of Wisconsin–Madison. Launched in the fall of 2020, their data science major became the fastest growing on campus, with 425 students declaring the major within just eight months [4]. This trend underscores the rising interest among students in acquiring data science expertise.

The proliferation of data science programs aligns with the labor market's demand for data scientists. The U.S. Bureau of Labor Statistics projects a 36% growth rate for data science jobs between 2021 and 2031, significantly outpacing the average for all occupations [6]

Over the past decade, data science has emerged as a pivotal field, leading numerous universities to establish comprehensive undergraduate programs integrating computer science, statistics, and domain expertise, allowing students to specialize in areas such as business analytics, computational biology, or social sciences. These

programs are designed to provide students with a strong foundation in data analysis, statistics, and machine learning. They emphasize both theoretical and practical aspects of data science and prepare students to handle real-world data challenges [5].

KEY DIFFERENCES BETWEEN CLINICAL DATA MANAGEMENT AND CLINICAL DATA SCIENCE

Aspect	CDM (Clinical Data Management)	CDS (Clinical Data Science)
Primary Goal	Maintain data accuracy and consistency	Ensure data quality and relevance
Quality Method	Rely on routine quality checks	Embed quality within the design phase
Thought Process	Emphasizes structured, logical outcomes	Prioritizes analytical and critical thinking
Trial Type Focus	Traditional randomized controlled trials	Flexible, adaptive, and master protocols
Key Data Source	Data collected directly from study sites	Data from electronic sources and decentralized trials (DCTs)
Process Approach	Standardized processes across multiple studies	Custom, risk-based processes tailored to each study
Data Volume	Manages relatively low data volumes from fewer sources	Handles large, complex data from multiple sources
Complexity of Data Flows	Simple and straightforward data pipelines	Complicated data interactions and flows
Vendor Role	Oversee basic vendor management tasks	Provides oversight and collaboration with vendors
Data Handling Tasks	Focus on cleaning and correcting data	Review, curate, tag, and manage data exclusions
Leadership Focus	Project-oriented leadership	Cross-functional collaboration and leadership
Standards Used	Follows clinical research standards	Integrates both clinical research and broader healthcare standards
Data Scope	Restricted to clinical trial datasets	Encompasses clinical trial and healthcare data sources
Programming Tools	Utilizes conventional languages like SQL, C#, and SAS	Leverages machine learning and AI tools like Python and R
Data Querying	Basic querying using traditional databases (e.g., SQL)	Advanced analysis using non-SQL, AI, and machine learning techniques

PREPARE YOURSELF FOR SUCCESS IN THE NEW ENVIRONMENT

Transitioning from a traditional Clinical Data Manager role to a Clinical Data Scientist role requires a proactive approach to skill development and career growth. Here are some strategies to help you succeed in this evolving landscape.

DEVELOP TECHNICAL SKILLS

Gain proficiency in programming languages such Python, R, or SQL. These are essential for analyzing and managing complex datasets. Whereas Clinical Data Managers of the past have been provided static versions of reports for data cleaning and action, the current expectation is for Clinical Data Scientists to actively interact with live clinical data to monitor trends and perform data cleaning functions. This requires basic competency in these technical skills.

Learn the basics so you understand advanced analytics. Many people talk about AI and machine learning. Make sure you are prepared to engage in the discussion and understand what activities are included in these topics. There are several online options for learning the basics of machine learning, AI, and predictive modeling.

There are many free sources to learn a little more about these programming resources. First consult your internal company resources to see what is offered. If there are no internal options available at your organization, look at the following free options. Be sure to review the terms and conditions and well as privacy agreements before signing up.

- DataCamp (https://www.datacamp.com/)

  Both R and Python tutorials are available, focusing on data manipulation, visualization, and statistics. While much is paid, several introductory courses are free.

- Kaggle (https://www.kaggle.com/)

  A data science community offering free tutorials on Python and R. Users can access real datasets and compete in challenges.

- Codecademy (Free Version) (https://www.codecademy.com/)

  Offers introductory Python courses to learn the basics

In addition, there are several Massive Open Online Courses available.

- edX (https://www.edx.org/)

  Offers free courses from universities like Harvard and MIT.

  - Suggested courses:
    - *Data Science: R Basics* (Harvard)
    - *Python for Data Science* (IBM)
    - *Introduction to predictive models, neural networks, and AI techniques.*

- Coursera (Free Audit Option) (https://www.coursera.org/)

  Free access to many courses when you audit.

  - Suggested courses:
    - *R Programming* (Johns Hopkins University)
    - *Python for Everybody* (University of Michigan)
    - *Machine Learning* by Stanford University (Andrew Ng)
    - *AI for Medicine Specialization* by DeepLearning.AI (clinical focus)

Also, consider collaborating on Open Source Projects on GitHub. There are many Python and R projects where you can gain real world experience.

- Open Source Projects on GitHub – Collaborate on Python or R projects to gain real-world experience.

SEEK OUT OPPORTUNITIES FOR CHANGE AND DEVELOPMENT

A key step in ensuring success in this constantly changing environment is the development of a growth mindset, embracing challenges as opportunities to learn and continuously improve your technical and analytical capabilities.

Stay Current with Industry Trends:

- Regularly engage with industry publications, webinars, and conferences to keep up with emerging technologies and methodologies.
- Subscribe to newsletters or join professional organizations like SCDM to stay informed.
- Use this knowledge to identify and implement best practices in your work.

Collaborate Across Functions:

- Strengthen communication skills to effectively collaborate with biostatisticians, programmers, and regulatory experts.
- Participate in cross-functional meetings to understand the broader goals of clinical trials and how data supports them.
- Build relationships that facilitate knowledge sharing and innovation within your organization.

Seek Mentorship and Networking Opportunities:

- Connect with professionals who have transitioned into Data Science roles to gain insights into their journey.
- Join online forums or attend industry events to network with peers and experts.
- Actively seek feedback and mentorship to identify areas for improvement and guide your career development.

Stay Open to Change

- Approach new challenges as learning opportunities and remain open to adapting to changes in technology and processes.
- Regularly reflect on your progress and set achievable goals to maintain motivation and focus.
- Embrace continuous learning to keep your skills relevant and impactful in the evolving landscape of clinical trials.

Become a Clinical Data Science Champion within your Organization:

- As you develop these skills share them within your organization.
- Become a change agent.
- Train others and serve as a mentor.

By adopting these strategies, you can position yourself as a valuable contributor to the future of clinical trials.

CONCLUSION

The role of the Clinical Data Manager is evolving, driven by advancements in technology, the diversification of clinical data sources, and the increasing complexity of clinical trials. This transformation offers Clinical Data Managers the opportunity to step into Clinical Data Science roles, where they can use advanced analytics to extract insights and drive trial success. By embracing new technologies, mastering data science principles, and fostering collaboration, Data Managers can navigate this transition confidently and play a pivotal role in the future of clinical research. By looking for opportunities to learn new skills and serving as a change agent in your organization, you will be set up for success in this ever changing environment.

REFERENCES

[1] The Evolution of Clinical Data Management into Clinical Data Science. https://scdm.org/wp-content/uploads/2024/03/SCDM-Position-Paper-Evolution-into-Clinical-to-Data-Science-V9.0.pdf

[2] ICH E6 (R3) Guideline for good clinical practice (GSP). https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e6-r3-guideline-good-clinical-practice-gcp-step-5_en.pdf

[3] The Rise of Data Science and Data Analytics Programs. https://www.encoura.org/resources/wake-up-call/the-rise-of-data-science-and-data-analytics-programs/

[4] New Data Science Degree Emerges as the Fastest Growing Major at UW-Madison. https://cdis.wisc.edu/new-data-science-degree-emerges-as-the-fastest-growing-major-at-uw-madison/

[5] Best Undergraduate Data Science Programs. https://www.usnews.com/best-colleges/rankings/computer-science/data-analytics-science?_sort=rank&_sortDirection=asc

[6] Occupational Outlook Handbook : Data Scientists. https://www.bls.gov/ooh/math/data-scientists.htm.

ACKNOWLEDGMENTS

Special thanks to Amy Booth, Senior Programmer at Veramed for sharing her input and review of this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:
Author Name: Debra Anderson
Company: Veramed
Email: debra.anderson@veramedinc.com
Website: https://www.veramedinc.com/