phuse·connect
connect·share·advance
The Clinical Data Science Conference
APAC 2026
19–21 February
Novotel Hyderabad International Convention Centre

Paper PD05

Next-Gen Statistical Programming in Clinical Trials: Driving Faster, Reliable, and Smarter Outcomes

Mrityunjay Kumar, Ephicacy Lifescience Analytics, Bengaluru, India
Shashikant Kumar, Ephicacy Consulting Group, New Jersey, USA

Abstract

Statistical programming in clinical trials has long been a cornerstone in determining the safety and efficacy of new treatments. Next-gen statistical programming techniques are reshaping the future of clinical trials, making them faster, more reliable, and smarter. The convergence of AI/ML, cloud computing, advanced statistical models, and real-time decision-making tools is driving innovation across the entire clinical trial process.

As these tools and methodologies continue to evolve, they will lead to even more efficient, accurate decision-making, optimizing data analysis processes, accelerating the delivery of life-saving treatments and therapies and improving overall outcomes for patients.

This session explores how next-generation statistical programming methodologies built on automation, modular code, and open-source tools R, Python alongside SAS is enabling faster and better outcomes. Implementing these innovations through governance models, validation frameworks, and upskilling strategies like empowering teams to transition from traditional methods to scalable, intelligent programming ecosystems would benefit the industry.

Introduction

Clinical trials are becoming increasingly complex and data intensive. The clinical research landscape is evolving rapidly, driven by the need for faster drug development, more reliable analytics, and smarter use of diverse data sources. Traditional statistical programming approaches - primarily manual, slow and code-intensive are often challenged by the increasing complexity of modern clinical trials, multi-modal data integration, and stringent regulatory expectations. Next-generation (Next-Gen) statistical programming introduces automation, artificial intelligence (AI), metadata-driven design, and reproducible workflows to accelerate insights while maintaining compliance and quality.

Current Trends in Statistical Programming

Current programming workflows are typically linear and dependent on manual coding and review cycles.

Common challenges may include:

- Manual coding errors and inefficiencies
- Difficulty in maintaining reproducibility and traceability
- Long turnaround times for data derivations and analyses
- Lack of standardized metadata and integration across systems
- Challenges in handling high-dimensional and real-world data
- Resource limitations and skill gaps within programming teams

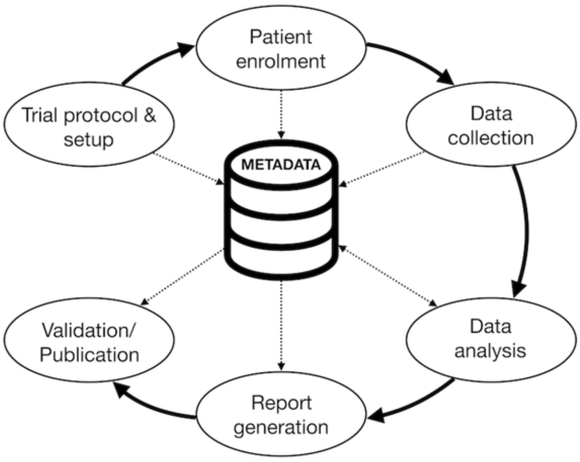


Figure 1: A conceptual diagram illustrating metadata-driven clinical trial lifecycle

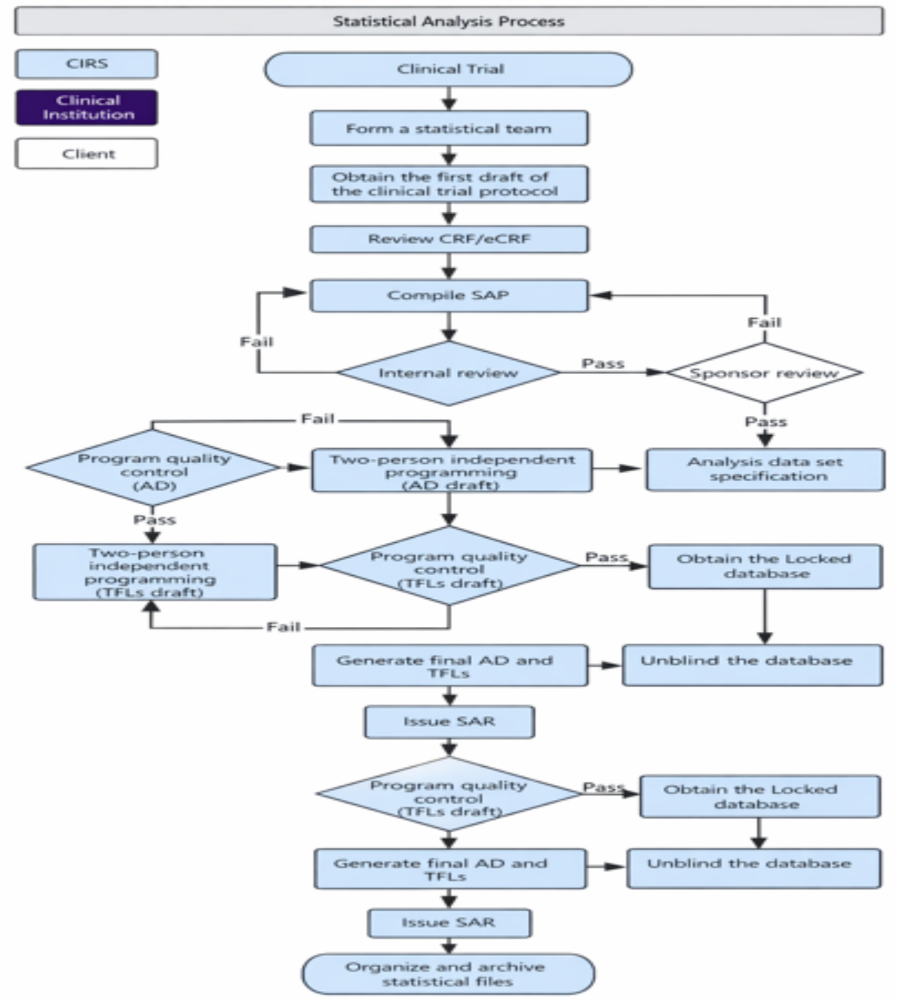


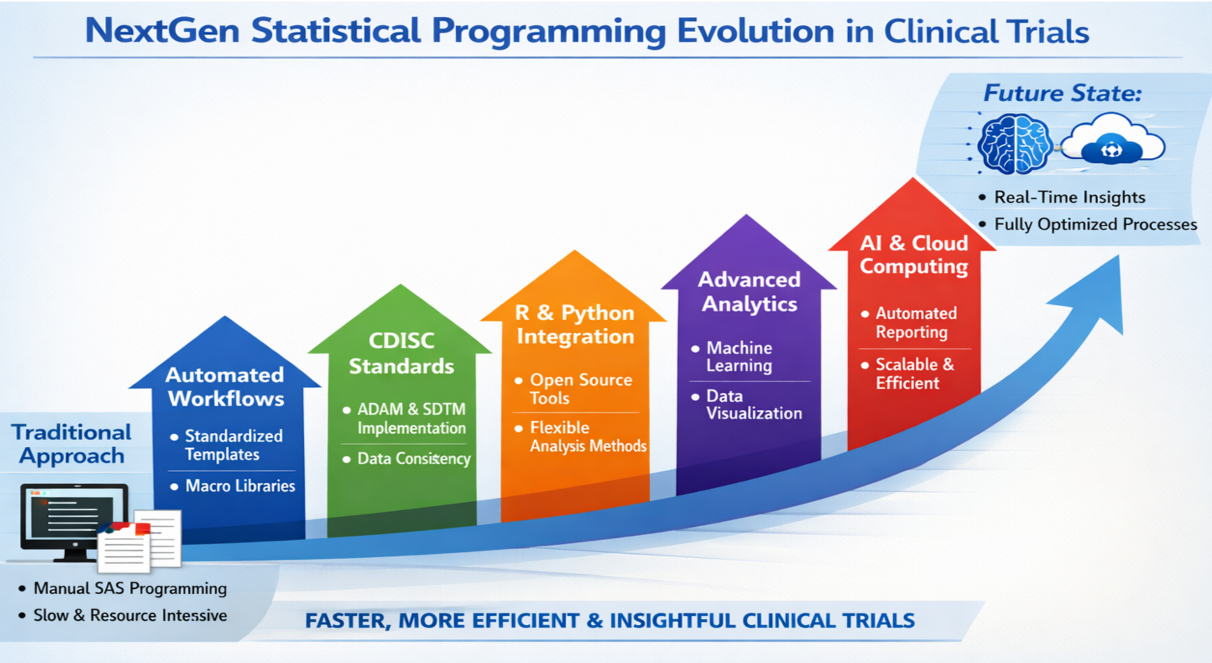
Figure 2: A Flowchart illustrating a standard clinical trial statistical analysis process

What Does “Next-Gen Statistical Programming” Mean?

Next-Gen statistical programming leverages modular, automated, and metadata-driven approaches that transform how data is processed, validated, and analysed using SAS and open-source platforms such as R, Python, GenAI etc. It focuses on enhancing productivity, ensuring data traceability, and enabling intelligent automation through AI and machine learning (ML).

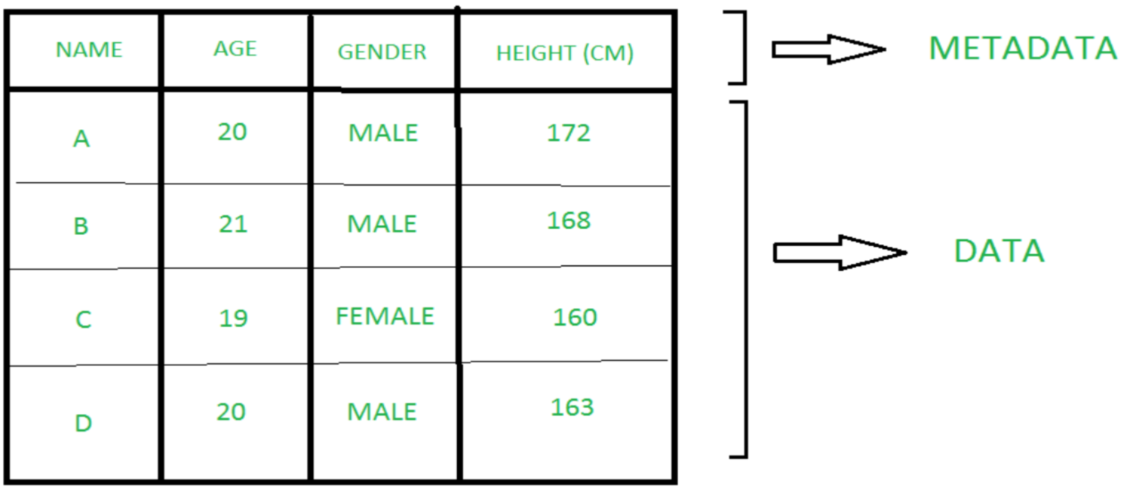
Core principles include:
• Automation and reproducibility
• Metadata and standards-driven designs
• Integration of AI/ML for decision support
• Continuous validation and audit trails
• Scalable, cloud-based computing environments

Figure: Evolution of Statistical Programming



Key Components of Next-Gen Statistical Programming

1.1 Metadata-Driven Frameworks: Centralized metadata defines variables, derivations, and data flow logic, enabling auto-code generation and consistent documentation.



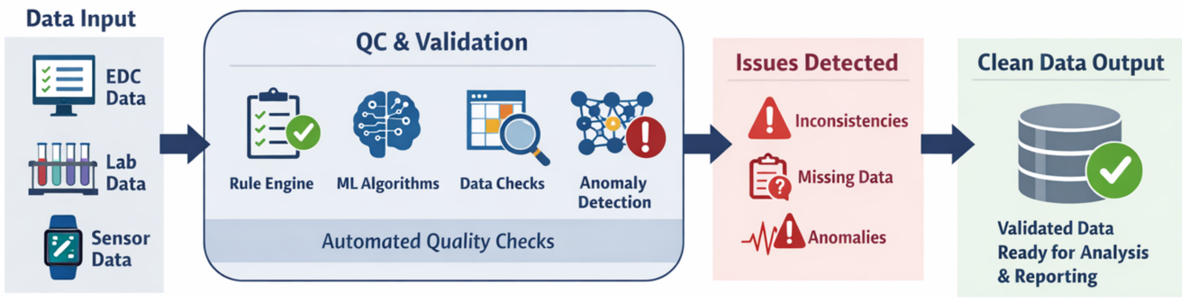
NAME	AGE	GENDER	HEIGHT (CM)
A	20	MALE	172
B	21	MALE	168
C	19	FEMALE	160
D	20	MALE	163

Figure 1: A simplified structure of data and metadata

1.2 Automated QC and Validation: Automated Quality Control (QC) and validation represent a transformative shift in the way clinical data is checked, verified, and audited. Traditionally, QC relied heavily on manual double programming, visual inspections, and independent review practices that are time-consuming, error-prone, and difficult to scale.

With automation, QC becomes systematic, reproducible, and integrated within the analytical workflow. Automated QC ensures **early detection of inconsistencies**, minimizes human error, improves regulatory compliance, and accelerates the delivery of submission-ready TFLs and datasets.

Built-in rule engines and ML algorithms identify inconsistencies, missing values, and anomalies earlier in the process.



1.3 Modular Code Reuse: Template-driven macros and reusable components are developed once and leveraged across multiple studies with minimal modification which accelerate programming and ensure uniformity across studies.

Instead of rewriting code for each clinical trial, programmers configure **study-specific metadata** (e.g., treatment arms, endpoints, visit structure), which dynamically drives standardized macros and functions.

This enables:

- ✓ Rapid study setup
- ✓ Consistent outputs across trials
- ✓ Reduced programming errors
- ✓ Improved maintainability and scalability

1.4 AI/ML Integration: Seamless use of **SAS**, **R**, and **Python** supports advanced statistical methods, data visualization, and exploratory analytics while maintaining regulatory rigor. Predictive modelling, natural language processing, and anomaly detection enhance efficiency and intelligence

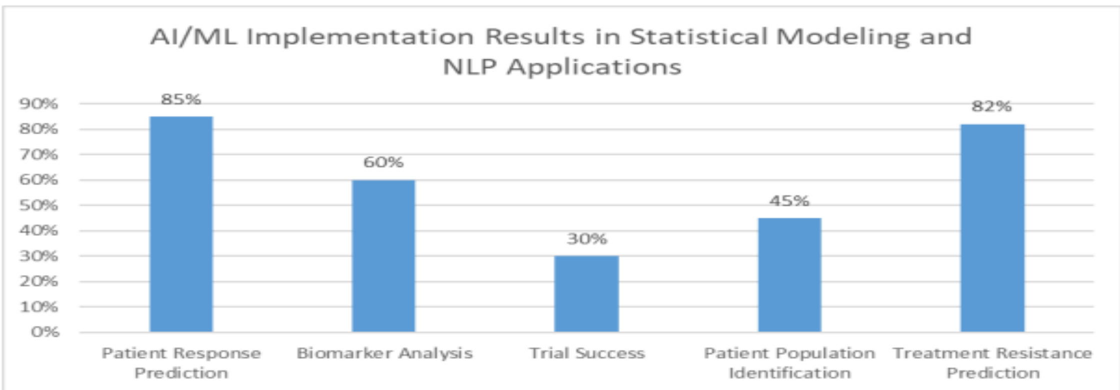


Figure 2: Quantitative Analysis of AI Performance in Clinical Trial Analytics and NLP Applications

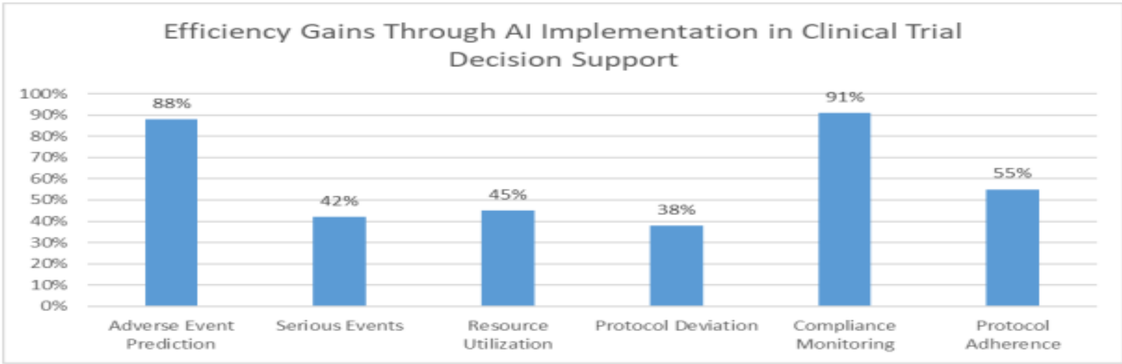


Figure 3: Quantitative Impact Analysis of AI-Driven Real-Time Processing and Adaptive Trial Management

1.5 Scalable Cloud Infrastructure: Containerized and parallelized environments support faster execution and collaboration. For example - a healthcare application can use a private cloud for processing sensitive data for compliance and a public cloud for data analysis to achieve cost-efficiency and scalability. This architecture enables a flexible and scalable approach, allowing efficient application deployment and operations in a multi-cloud setting, while maintaining consistency and resilience across platforms.

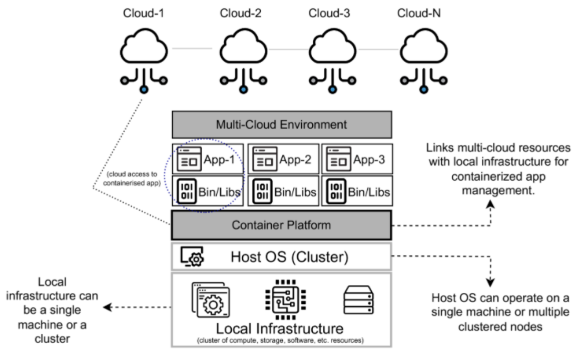
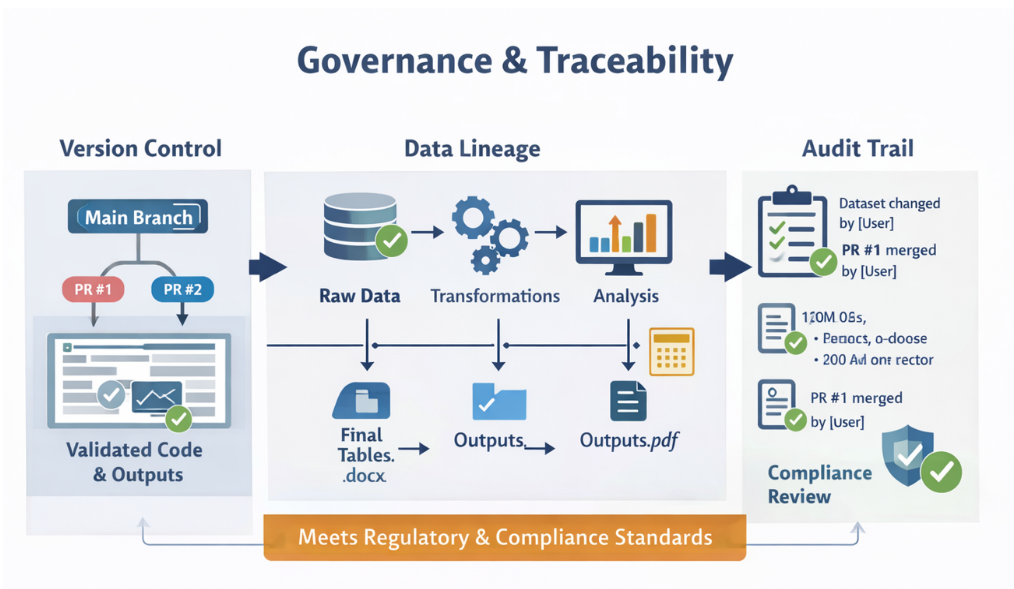


Figure 4: Containerization in multi-cloud environment [: Ref1]

1.6 Governance and Traceability: Complete audit trails, version control, and data lineage tracking meet compliance and regulatory standards. It provides Automated Logging, timestamps, versions control and support for regulatory audit requirements.



Version Control – Version control refers to systematic tracking of changes made to programs, datasets, specifications, macros, and outputs throughout the clinical data lifecycle. It includes capturing "who changed what, when, and why" using structured tools and processes.

Data Lineage – Data lineage describes the complete flow of data from source acquisition through transformation, derivation, integration, and final analysis outputs. It visually and programmatically shows how each variable and dataset was derived.

Components of Data Lineage:

- **Source Data:** EDC, ePRO, labs, imaging, wearables
- **Transformation Steps:** Data cleaning, imputation, derivations, business rules
- **Intermediate Datasets:** SDTM domains, supplemental qualifiers, mapping tables
- **Analysis Datasets:** ADaM structured datasets for efficacy, safety, PK/PD
- **Outputs:** TFLs, submission-ready datasets, traceability documents

Audit Trail – An audit trail is an immutable, chronological log of all activities related to data processing, programming, and analysis. It records every action applied to data, including creation, modification, deletion, access, approvals, and system-level events.

Component	Purpose	Benefits
Version Control	Tracks code and data changes	Reproducibility & collaboration
Data Lineage	Shows transformation paths	Transparency & validation
Audit Trail	Logs who did what & when	Regulatory compliance

Traditional vs Next-Gen Reporting

Key Metric	Traditional Process	Modernized Approach
Programming Cycle Time	8–10 weeks	3–4 weeks
QC Rework	High (manual)	Low (automated)
Regulatory Compliance	Reactive	Continuous
Collaboration	Siloed	Integrated
Data Insight Speed	Weeks	Near Real-Time

Table 2: Performance Analysis of AI-Driven CDISC Standards and Regulatory Documentation (https://iaeme.com – IAEME Publication)

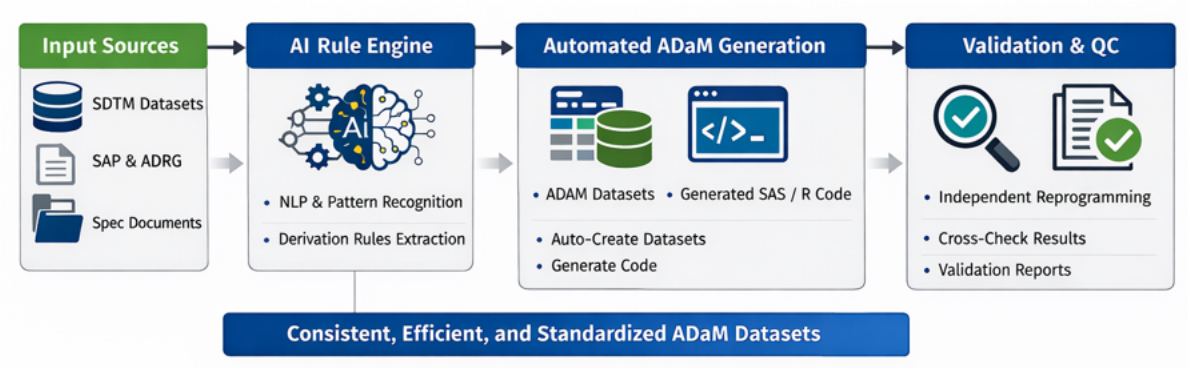
Compliance Category	Performance Metric	Achievement Rate
Data Mapping	CDISC Standards Accuracy	91%
Variable Processing	Simultaneous Processing Capacity	500,000
Dataset Validation	Time Reduction	65%

Compliance Category	Performance Metric	Achievement Rate
Error Detection	Improvement Rate	58%
Regulatory Documentation	Time Reduction	70%
Compliance Issues	Detection Accuracy	94%
Protocol Analysis	NLP Processing Accuracy	89%
Quality Control	Documentation Error Reduction	45%

Real Time Examples:

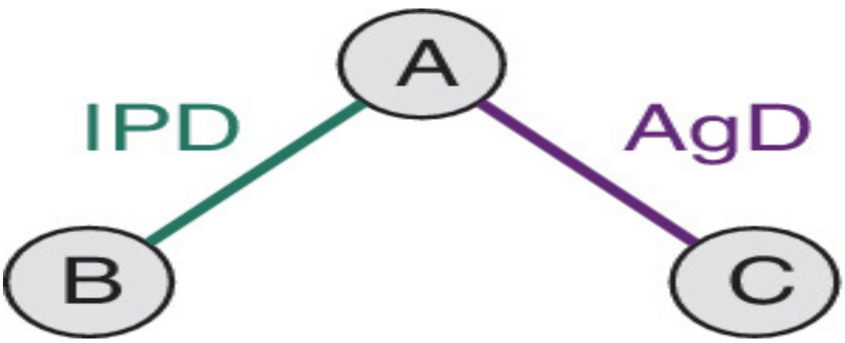
Example 1: Automated ADaM Derivation and QC

NLP models read SAP, ADRG, and specifications and AI extracts derivation rules (e.g., how AVAL, BASE, CHG were historically computed). Automated pipelines generate ADaM datasets and validate with independent programming (SAS/R). Reduces manual effort in programming ADaMs by 40–60%, ensures consistency across therapeutic areas (TAs) and helps standardize datasets across CROs/vendors.



Example 2: AI assisted MAIC (Matching-Adjusted Indirect Comparison) Analysis

It makes indirect treatment comparisons more reliable when head-to-head data are unavailable. It is widely used in health technology assessments when regulatory decisions depend on comparative effectiveness evidence across different trials.



Challenges, Risks, and Mitigation Strategies

While Next-Gen programming offers tremendous benefits, implementation is not without challenges:
• Regulatory validation and explainability of AI components
• Integration with legacy systems and diverse data formats
• Change management and workforce training
• Ensuring privacy, data security, and governance

Mitigation strategies include adopting hybrid models, establishing robust validation frameworks, maintaining human oversight, and investing in continuous learning.

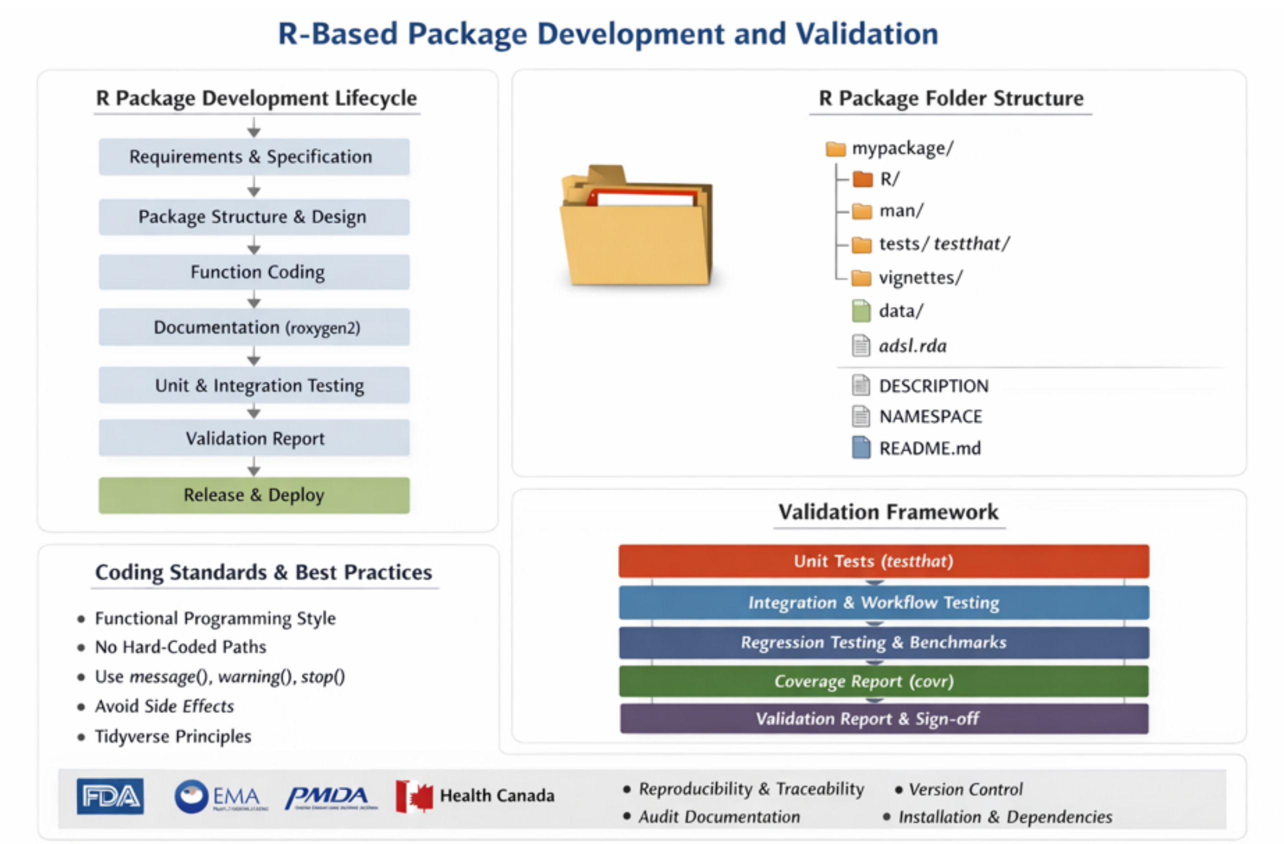
Future Trends and Directions

The next decade will see deeper integration of generative AI for auto-coding and documentation, federated learning for privacy-preserving collaboration, and real-time analytics for adaptive designs. Regulatory frameworks are also evolving to embrace explainable AI and standardized metadata for machine-driven processes. Future tools will allow 'citizen programmers' to perform analytics through intuitive interfaces, expanding accessibility while maintaining rigor.

R-Based Package Development and Validation

Developing and validating an R package for clinical or statistical programming requires a structured, regulatory-aligned framework. It ensures transparency, reproducibility, traceability, and long-term maintainability across clinical trials. A well-designed R package not only standardizes code but also accelerates quality control, supports automated checks, and integrates seamlessly with CDISC data structures (SDTM, ADaM).

Validation ensures the package is robust, reliable, and compliant with industry expectations such as FDA, EMA, PMDA, and Health Canada.



Conclusion

Next-Gen statistical programming represents a paradigm shift in clinical research analytics. By leveraging automation, metadata, and AI, organizations can achieve faster insights, improved quality, and more efficient regulatory submissions. A gradual, well-governed adoption strategy ensures successful integration of these innovations into existing workflows - paving the way for smarter, data-driven clinical development.

The future of clinical trials lies in smarter, reproducible, and integrated statistical programming. Organizations must embrace innovation to remain agile and regulatory-ready. The journey begins with upgrading skills, systems, and mindset.

Acknowledgments

Authors would like to extend their sincere thanks to Ephicacy Lifescience Analytics for giving them an opportunity to write and present this paper. Any brand and product names are trademarks of their respective companies.

Recommended Reading

Lex Jansen. (2025). Leveraging AI/ML in Statistical Programming: Enhancing Efficiency, Compliance, and Insights in Clinical Trials.

https://iaeme.com/MasterAdmin/Journal_uploads/IJRCAIT/VOLUME_8_ISSUE_1/IJRCAIT_08_01_097.pdf

https://arxiv.org/html/2403.12980v2

Cytel Perspectives. (2025). Streamlining Data Management and Improving Statistical Accuracy in Clinical Trials with AI.

PHUSE Whitepaper. (2024). Future of Statistical Programming: Automation and AI Integration.

EAJournals. (2025). Automation in Statistical Programming: Advancing Clinical Research Through AI Integration.

FDA. (2023). Good Machine Learning Practice (GMLP) for Medical Devices – Guiding Principles.

FDA Study Data Technical Conformance Guide (2023).

FDA 21 CFR Part 11 – Electronic Records and Signatures.

Assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study: David M. Phillippo, Sofia Dias, A. E. Ades, Nicky J. Welton https://doi.org/10.1002/sim.8759

Contact Information

Your comments and questions are valued and encouraged. Contact the authors at:

Mrityunjay Kumar, Associate Director II - Statistical Programming
Ephicacy Lifescience Analytics
Bengaluru, India
E-mail: mrityunjay.kumar@ephicacy.com
www.ephicacy.com

Shashikant Kumar, Director - Statistical Programming
Ephicacy Consulting Group
New Jersey, USA
E-mail: shashikant.kumar@ephicacy.com
www.ephicacy.com