



SPA MDR CDR SCE

AS14

Big Data Analytics: Unlocking Insights from Vast Clinical Datasets

MAR 2025



The Clinical Data
Science Conference

US 2025

16–19 March
Renaissance Orlando at SeaWorld Florida



About the Presenters



Kala Shiva

Executive Director
of Services



kala.shivali@sycamoreinformatics.com



Cary,
North Carolina
USA



<https://www.linkedin.com/in/chandrakala-shivalingaiah-443b333/>



Bio

Kala has over 30 years of experience across BioPharma, Biotech, CRO, CRM, customer service, finance, and IT. She has held key roles in the pharmaceutical sector, including liaison, acting head of SP, and program director. She holds a bachelor's in computer science engineering from Bangalore University and an MBA in Bioscience from NC State University. Currently, she leads the Service team at a SaaS company specializing in BioPharma and Biotech.

Her expertise lies in understanding business needs and driving innovation for future growth by harnessing modern technology. Outside of work, Kala enjoys reading, podcasts, marathons, hiking, and long walks. She is excited to contribute to and learn from this community.



Bill Qubeck

Senior Director of
Services



bill.qubeck@sycamoreinformatics.com



Apex,
North Carolina
USA



<https://www.linkedin.com/in/williamqubeck/>



Bio

With nearly 30 years in the pharmaceutical industry, Bill has led large-scale digital transformations, optimizing R&D and IT operations to streamline clinical and biometrics functions. He has worked with nearly all the top 50 pharmaceutical companies and CROs, driving measurable business impact. His expertise lies in innovation, operational efficiency, and leveraging technology to maximize value across the drug development lifecycle.



Agenda

- Introduction
- Big Data Sources & Challenges
- Cloud Computing Benefits & Considerations
- Interoperability & Standards
- Integration with Statistical Computing Environment (SCE)
- Case Studies
- Regulatory & Ethical Considerations
- Summary and Q & A



Introduction

- Big Data in clinical research represents one of the most significant advancements in pharmaceutical innovation; clinical trials are now analyzing vast & diverse datasets
- Sourced from electronic health records, wearable devices, genomic sequencing, real-world evidence, and medical imaging are enabling researchers to identify patterns, optimize trial design, and accelerate drug discovery
- Integration of AI & ML in clinical trials introduced predictive modeling for patient recruitment, anomaly detection in trial data, & AI-assisted regulatory submissions
- Ability to analyze structured and unstructured data in real time is transforming decision-making, reducing trial durations, and improving patient outcomes. However, the increased reliance on data-driven methodologies also introduces challenges related to data standardization, regulatory compliance, security, and ethical concerns



Sources of Big Data in Clinical Trials

01

Electronic Health
Records (EHRs)

02

Wearable Devices
& Remote
Monitoring

03

Omics Data
(Genomics,
Proteomics, etc.)

04

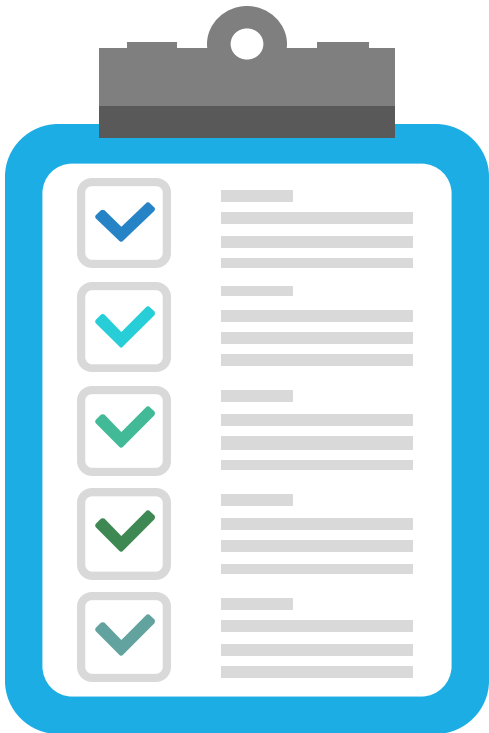
Real-World
Evidence (RWE)
and Claims Data

05

Imaging and
Multi-Modal
Data



Challenges of Managing and Analyzing Large Datasets



1

Data Integration and Standardization

- Clinical trial data has many sources, each using different formats and terminologies. Integrating un/structured data remains a major hurdle

2

Data Privacy and Compliance

- Strict adherence to regulations like GDPR
- Data de-identification, encryption, and controlled access

3

Computational and Storage Challenges

- Requires high-performance computing (HPC) and scalable cloud platforms

4

Bias and Data Quality Issues

- EHR, RWE, etc. data may not fully capture diverse patient populations, leading to biased outcomes

5

Ethical and AI Interpretability Concerns

- Algorithm transparency, explainability, etc. (e.g., black-box AI models), pose questions and challenges



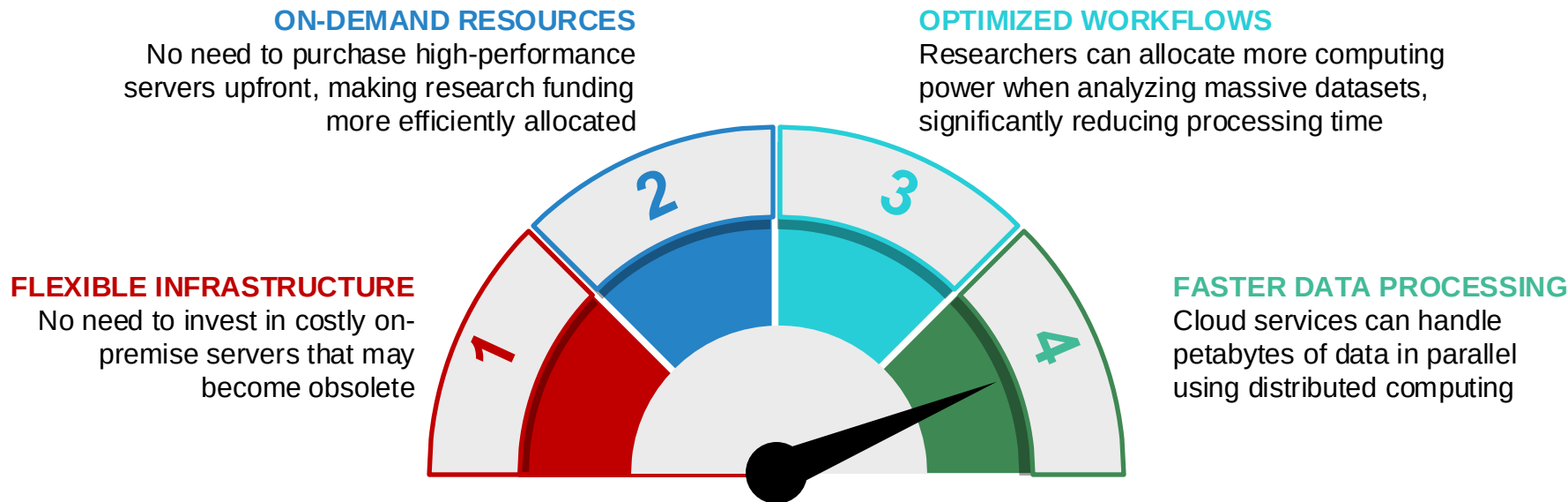
Analytical Tools



- SAS remains the dominant programming language in clinical trials and regulatory submissions
- R has been rapidly gaining traction in biomedical research due to its open-source nature, powerful statistical capabilities, and widespread adoption in biostatistics and bioinformatics
- Python has also emerged as a major player in biomedical research, driven by its flexibility, cost-effectiveness, and strong capabilities in artificial intelligence (AI) and machine learning (ML)
- SQL plays a crucial role in managing clinical and biomedical databases, as it is the primary tool for querying and extracting patient data from electronic health records (EHRs) and clinical data warehouses



Cloud Computing for Scalable Data Processing



One of the biggest advantages of cloud-based analytics is its ability to scale resources on demand. Clinical trials and biomedical research generate enormous datasets, including genomic data, electronic health records (EHRs), imaging files (MRI, CT scans), and real-time patient monitoring data. Cloud platforms, such as AWS, Microsoft Azure, and Google Cloud, provide elastic computing power, ensuring that research teams can analyze large datasets without the limitations of on-premise infrastructure.



Cloud Security and Compliance Challenges



1

DATA PRIVACY

HIPAA and GDPR. Strict requirements for handling, storing, & transmitting sensitive health data. Must ensure compliant storage solutions, encryption protocols, and secure access controls to protect patient confidentiality.

2

CYBERSECURITY RISKS

Cloud-based storage increases vulnerability to cyberattacks, data breaches, and unauthorized access. Essential to implement encryption standards, firewalls, intrusion detection systems, & multi-factor authentication (MFA)

3

DATA RESIDENCY LAWS

Restrictions on where data can be stored and processed. Many countries have regulations requiring that patient data remain within national borders. Carefully evaluate data sovereignty implications and choose cloud architectures that align with local requirements



Interoperability Challenges

Lack of Standards

- Patient data is not stored within same technology, field attributes, etc. across healthcare institutions and research organizations
- This lack of standardization increases the difficulty and effort to aggregate and analyze the data

Terminology & Coding

- Medical data is coded using multiple standards, including: ICD-10, SNOMED CT, LOINC
- Coding differences create barriers to data integration and cross-platform analytics
- Without harmonization, patient records may lack consistency, leading to errors in processing, results, etc.

Security & Compliance

- Interoperability to align with regulatory rules like HIPAA, GDPR, and FDA 21 CFR Part 11, which govern data security, patient privacy, and electronic records
- Sharing clinical trial data across international research institutions requires secure data exchange protocols and access control mechanisms



No One Common Standard: FHIR, CDISC, & OMOP - Definitions

- FHIR (Fast Healthcare Interoperability Resources) is a widely adopted standard developed by HL7 (Health Level Seven International) that enables:
 - Standardized APIs for EHR integration, making data exchange between hospitals and research institutions more efficient
 - Data representation in structured formats (JSON, XML, RDF) for interoperability
- CDISC (Clinical Data Interchange Standards Consortium) defines structured data models for regulatory submissions (e.g.,)
 - SDTM (Study Data Tabulation Model): Ensures uniform data collection for clinical trials.
 - ADaM (Analysis Data Model): Facilitates statistical analysis and regulatory review.
- OMOP Common Data Model open-source framework to standardize real-world data (RWD) and observational health datasets EHRs, insurance claims, and registries
 - OMOP plays a crucial role in post-market surveillance, epidemiological studies, and comparative effectiveness research; facilitates research on patient outcomes over time
 - Multi-database research: Allows data harmonization across different healthcare systems globally



FHIR, CDISC, & OMOP Standards - Focus

Feature	FHIR	CDISC (SDTM/ADaM)	OMOP CDM
Primary Use Case	Clinical data exchange	Regulatory clinical trial submissions	Observational and real-world data research
Data Sources	EHRs, medical devices	Clinical trial datasets	EHRs, insurance claims, registries
Regulatory Requirement	Not mandatory	Required by FDA, EMA, PMDA	Not mandatory but widely used in RWE studies
Flexibility	High (designed for API-based integration)	Structured for standardizing clinical trial data	High (used for multi-database research)



Integration with Statistical Computing Environment (SCE) for Big Data



Consume metadata standards including transformations and mappings



Standardize handling of data (collection to locking to archiving)



Data ingest: Handle data from any source, in any format



Define and manage role-based access across the lifecycle



Reliably audit access and changes



Coordinate tasks, track work and enforce processes



Maintain traceability and lineage across programs, data and outputs with version control



High-performance computing environment that is scalable to support the access, processing, analysis, and storage of large data



Connect to data sources, up & downstream applications via integrations and REST APIs



Case Study 1: Predictive Modeling for Patient Recruitment

- Up to 80% of clinical trial delays = challenges in enrolling eligible participants
- Traditional recruitment methods (manual screening of EHR & physicians' referrals) are time-consuming & often ineffective in meeting enrollment targets
- Machine Learning algorithms can analyze EHRs and RWE data to efficiently screen and match patients. AI models can predict the likelihood of patient dropout by assessing historical health data, socioeconomic factors, and treatment adherence patterns
- Predictive modeling can patient diversity by incorporating demographic and social determinants of health (SDOH) data, ensuring broader representation in clinical research



Case Study 1: Predictive Modeling for Patient Recruitment, Continued

- IBM Watson Health's initiative with Mayo Clinic: AI-powered
 - AI-driven system analyzed over one million patient records in real-time
 - Researchers were able to match potential candidates with trials based on genomic data, demographic characteristics, and medical history
 - It demonstrated a significant increase patient enrollment rates because of this IA-driven eligibility screening process
 - AI Natural Language Processing (NLP) extracted relevant trial criteria from unstructured clinical notes, significantly reducing manual workload and improving the accuracy of patient selection



Regulatory and Ethical Considerations

AI-DRIVEN ANALYTICS

Data bias, algorithmic transparency, & informed consent
Raises ethical questions about fairness, accountability, and unintended consequences

AI ALGORITHMIC BIAS

AI models trained on non-representative datasets can amplify existing biases
AI frameworks must prioritize bias detection and mitigation strategies

1

2

DATA PRIVACY AND COMPLIANCE

Solutions must be designed to protect personal health information (PHI), ensure data security, and prevent misuse

PATIENTS RIGHTS

Define, protect, inform, and enforce patients rights as well as provide transparency and access

3

4



Summary

- Rapid evolution of Big Data analytics in clinical research has already demonstrated its profound impact on patient recruitment, clinical trial efficiency, regulatory compliance, and drug development timelines
- Researchers can now process complex datasets with unprecedented speed with power of cloud computing, real-time data integration, and AI
- Despite its numerous benefits, the integration of Big Data and AI into clinical research is not without challenges
- Data security, privacy regulations, interoperability concerns, and ethical considerations remain major questions that must be addressed
- Regulatory agencies such as the FDA, EMA, and European Commission have already begun introducing new guidelines on AI in drug development, privacy protection, and real-world evidence usage, but continued industry collaboration is necessary to refine these frameworks



Connect with the Presenters!

Kala Shiva



Bill Qubeck



Thank you & Q&A



Thank you!

