



TITLE: Statistical Evaluation of Probabilistic Generative AI Models

SPEAKER: Christian Schmid

ABSTRACT:

Background: The rapid integration of Generative AI (GenAI) in regulated industries, such as drug development, presents a paradigm shift for traditional validation practices. Unlike traditional GxP software, which is rigid and deterministic, GenAI is inherently probabilistic and prone to hallucinations. This necessitates a move from the "gold standard" of verifying absolute correctness toward a robust statistical framework for managing and quantifying uncertainty.

Objectives: This presentation outlines a comprehensive statistical framework for the validation of "Locked and Static" GenAI models—systems that do not adapt independently post-deployment—to ensure reproducibility and safety in GxP environments.

Methods & Framework: We explore three core pillars for improving and evaluating model reliability:

- **Fine-Tuning:** Applying classical machine learning evaluation to categorization tasks.
- **Retrieval-Augmented Generation (RAG):** Implementing specialized metrics to evaluate the "Retriever" (Context Precision and Context Recall) and the "Generator" (Faithfulness and Groundedness).
- **Prompt Engineering:** Developing structured system prompts evaluated through binary and Likert scales to measure adherence to complex instructions.

Conclusion: As regulatory bodies move toward human-centric, risk-based AI governance, statisticians must lead the transition from deterministic verification to uncertainty management. This framework provides the necessary statistical rigor to evaluate "Locked and Static" models, ensuring that generative applications in drug development meet the high standards of reproducibility and reliability required for GxP environments.

BRIEF SPEAKER BIO: Christian Schmid is a Senior Statistical Scientist at Roche and a founding member of the CMC Statistical Network Europe, where he has served on the steering committee since its inception. In his role, he primarily supports Roche's commercial manufacturing and has recently expanded his work to include several AI validation initiatives. He earned his PhD in Statistics from The Pennsylvania State University, with a specialization in computational methods.