33RD BIOPHARMACEUTICAL APPLIED STATISTICS SYMPOSIUM

AI for Biostatistics in Regulated Applications

Innovation, Scientific Judgment, and the Human Imperative

Zack Wintrob | Emmes

November 2–4, 2026 • Savannah, Georgia

AI opportunities across the biostatistical lifecycle

01 PROTOCOL + SAP

Study objectives, estimands, methods, power analysis

→

02 CRF + INTERIM ANALYSIS

Data feasibility, controlled reviews

→

03 SDTM MAPPING

Source semantics and traceability

04 ADaM + PROGRAMS

Derivation rules, analysis code

→

05 TFL + QC

Summaries, outputs, results, interpretation

→

06 CSR + TMF

Narrative, disclosure, records

QUALIFIED BIOSTATISTICIAN + CONTROL SYSTEMS: SCIENTIFIC OVERSIGHT AT EVERY HANDOFF

ACTION → Prioritize attention to downstream impact and data dependencies—not time savings alone.

Published AI performance is task-dependent—so is our experience

40%

less writing time

Noy & Zhang, 453 professionals

General writing tasks, not Clinical Trials

≈80%

protocol-writing rubric

Markey et al., with RAG

Clinical logic + reference scores + content relevance

19%

longer coding time

METR, experienced developers

2025 RCT; newer estimates uncertain

QUALITATIVE EMMES EXPERIENCE • 2 YEARS OF USE

ADaM | Mixed feedback

Some users report efficiency gains; others report substantial extra effort—especially for nonstandard derivations.

TFL-to-text | Favorable feedback

Users generally report easier drafting, but benefit and accuracy have not been reliably quantified.

ACTION → Evaluate net effort + final quality separately for each task class.

Sources: [NZ23] [MARK25] [METR25] [METR26]

A convincing AI output can still be scientifically wrong

FACTUAL

Invented data, variables, citations, or results

STATISTICAL

Wrong estimand, derivation rule, or inference

PROCEDURAL

Version control, operational sequence, or source trail

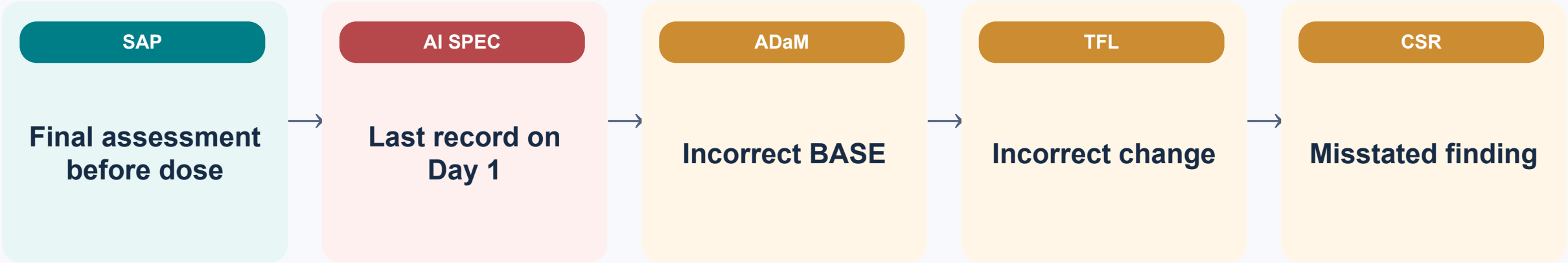
CONTEXTUAL

Ignored amendments, sponsor rules, or blinding

Fluency ≠ factual accuracy ≠ statistical validity ≠ regulatory readiness

ACTION → Test for consequential statistical failures—not only writing quality.

An upstream derivation error can spread into every deliverable



A structural conformance check may not discover an analytically incorrect baseline.

ACTION → Validate and version critical derivation rules at their origin.

Regulatory expectations depend on context of use

ICH E6(R3)

Reliability, risk-proportionate controls, computerized systems

FDA AI guidance

2025 DRAFT: credibility for AI-generated regulatory evidence

FDA–EMA principles

2026: human-centric design, multidisciplinary expertise

FDA study-data guide

Oct 2026: technical specifications for submission data

SCOPE NUANCE: Operational drafting without impact on safety, quality, or result reliability is outside the 2025 draft AI guidance.

ACTION → Specify intended use and impact before choosing controls.

Sources: [FDAAI25] [FDAEMA26] [E6R3] [FDASTD26]

Complexity and consequence determine when to bypass AI

LOWER IMPACT

Indexing, checklists

Verify completeness

MODERATE IMPACT

Specifications, narratives

Approved sources + independent review

HIGHER IMPACT

Interim decisions, core derivations

Enhanced controls or restricted use

MANUAL / EXPERT-AUTHORED ROUTE

Use when rare derivations cannot be reliably grounded or checked.

ACTION → "Do not use AI" is a valid risk-based workflow decision.

The critical question is not whether AI can write it

Can AI improve workflow efficiency without compromising scientific validity?

01 Bound the task

Reviewable outputs and approved context

02 Embed biostatisticians

Design, adaptation, test cases, adjudication

03 Prove fitness

Traceability, risk-based checks, total effort

ACTION → Put statistical expertise into tool design—not only final output review.

Protocol review: AI can challenge completeness

WORKFLOW APPLICATION

WHERE AI MAY HELP

Highlight missing or inconsistent endpoint, schedule, and methodological language.

ILLUSTRATIVE TASK

Rescue medication appears in the protocol, but handling as an intercurrent event is absent from the analysis strategy.

FAILURE MODE

AI may propose an inappropriate estimand strategy.

STATISTICIAN-LED CONTROL

Biostatistician defines the question, evaluates options, and approves design decisions.

ACTION → Use AI to challenge completeness, not choose trial design.

Sources: [SHIN26] [M11]

SAP development: drafting is not methodological judgment

WORKFLOW APPLICATION

WHERE AI MAY HELP

Draft sections and compare methods with approved protocol requirements.

ILLUSTRATIVE TASK

An MMRM section leaves covariance selection and missing-data sensitivity analyses unclear.

FAILURE MODE

Plausible prose may conflict with the target estimand.

STATISTICIAN-LED CONTROL

Biostatistician selects methods, assumptions, and robustness checks.

ACTION → Use AI to find gaps; the statistician owns analytical choices.

Sources: [E9R1]

CRF review: find analysis data gaps before enrollment

WORKFLOW APPLICATION

WHERE AI MAY HELP

Cross-check draft CRFs against planned endpoint and derivation requirements.

ILLUSTRATIVE TASK

A PK CRF records nominal sample time but not adequate actual collection and dosing timestamps.

FAILURE MODE

PK intervals cannot reliably be reconstructed (e.g. steady-state renal clearance is a listed endpoint – requires urine collection and plasma collection to align).

STATISTICIAN-LED CONTROL

Statisticians, data managers, and pharmacometricians confirm data feasibility and appropriateness.

ACTION → Review CRFs against the calculations they must eventually support.

Interim analysis: automation must protect trial integrity

WORKFLOW APPLICATION

WHERE AI MAY HELP

Check pre-specified output inventories, code, and report templates.

ILLUSTRATIVE TASK

AI flags an interim analysis table omitted from the pre-approved report package.

FAILURE MODE

Unblinding, altered stopping boundaries, or unauthorized results access.

STATISTICIAN-LED CONTROL

Respect firewall separation, pre-specified rules, and independent committee roles.

ACTION → Use AI for controlled execution—not trial continuation decisions.

Sources: [E6R3]

SDTM mapping: clinical meaning precedes syntax

WORKFLOW APPLICATION

WHERE AI MAY HELP

Propose source-to-SDTM mappings and identify incomplete metadata.

ILLUSTRATIVE TASK

Flag a mismatch between nominal and actual sample timing in exposure / PK mapping.

FAILURE MODE

A syntactically valid mapping may distort lineage or clinical semantics.

STATISTICIAN-LED CONTROL

Inspect source CRFs, applicable SDTMIG, terminology, and sponsor conventions.

ACTION → Apply deterministic conformance plus accountable semantic review.

Sources: [FDASTD26]

ADaM: a useful assistant for common rules, a burden for exceptions

CONTROLLED INPUTS

- SAP + protocol
- SDTM metadata
- Approved standards

→

AI PROPOSES

ADaM metadata + derivations

→

- REVIEW
- REWORK
- BYPASS AI

Observed user experience: routine rules may benefit; uncommon or study-specific derivations can substantially increase prompt and correction effort.

ACTION → Use AI selectively; use expert authorship when it is faster and safer.

Sources: [ADAM] [POPPK]

Synthetic ADaM example: one baseline rule, two answers

ADAPTED EXAMPLE • NOT EMMES DATA

ASSESSMENT	TIME	SCORE
Screening	Day −1	**18**
Pre-dose	Day 1, 07:55	**20**
First dose	Day 1, 08:00	**—**
Post-dose	Day 1, 08:30	**16**
Week 4	Day 28	**12**

INCORRECT AI-STYLE RULE

BASE = 16 → CHG = −4

APPROVED SAP RULE

BASE = 20 → CHG = −8

ACTION → Prove exceptional-case correctness; structural conformance is insufficient.

Sources: [ADAMCONF]

ADaM programming: running code is not the same as correct code

AI-assisted generation

Generate scaffolds and unit-test suggestions from an approved specification.

Fast for patterns; costly when study-specific rules require repeated repair.

Independent statistical verification

Test boundary times, missing dose records, ties, treatment changes, and analysis-population exceptions.

Do not accept tests that merely echo the generated code.

STOP RULE: When verification + prompting takes longer than expert authoring, bypass AI.

ACTION → Evaluate total author + correction + QC effort, not only code-generation speed.

Sources: [ADAMCONF]

TFL programming: use AI for context, code for reconciliation

AI-ASSISTED CONTEXT CHECKS

- Table title ↔ SAP alignment
- Footnotes and population definitions
- Missing or inconsistent shell content
- Discrepancy triage and drafting

DETERMINISTIC QC

- Values, denominators, rounding
- Group labels and ordering
- Analysis populations and missingness
- Reproducible checks against outputs

ACTION → Make numbers machine-verifiable and interpretation statistician-verifiable.

Synthetic TFL-to-text: fluency can reverse the finding

ADAPTED EXAMPLE • NOT EMMES OUTPUT

Source: Response table

Investigational	**9/60 (15.0%)**
Comparator	**6/60 (10.0%)**

FLAWED HYPOTHETICAL AI TEXT

"Statistically significant reduction in objective response rate with investigational treatment."

SOURCE-FAITHFUL TEXT

"In the investigational group 9/60 (15.0%) participants achieved objective response compared to 6/60 (10.0%) in the comparator group."

Two defects: wrong direction + unsupported significance. No inferential test is provided.

ACTION → Require source-cell reconciliation and human approval of all inferential claims.

CSR and TMF: AI can find gaps but cannot approve records

CSR

Verify consistency between text, analysis, and cited TFL versions.

CLINICALTRIALS.GOV

Prepare and QC structured summary results where reporting applies.

TMF

Flag missing versions, approval statuses, and record metadata.

AI checks completeness; designated personnel confirm the authoritative final record.

ACTION → Treat version, provenance, approval, and archive controls as human tasks.

Sources: [E6R3] [NIHCLIN]

Biostatisticians belong inside AI development, not at the end

01 SPECIFY

Statistical requirements, intended uses, prohibited decisions

02 DEVELOP

Gold examples, context, prompts, rules, knowledge sources

03 EVALUATE

Challenge datasets, severity of errors, expert adjudication

04 IMPROVE

Root cause, retraining/adaptation, change approval

STATISTICIAN DECISION RIGHTS THROUGHOUT THE LIFECYCLE

ACTION → Give statisticians authority over requirements, test sets, and release gates.

Sources: [FDAEMA26]

Measure total workflow performance—stratified by complexity

PROPOSED EVALUATION DESIGN

DEFINE
Intended use

→

CHALLENGE
Routine + nonstandard

→

ADJUDICATE
Expert expected result

→

DECIDE
Continue / restrict / improve

METRICS (EVALUATE TASKS SEPARATELY)

Critical statistical errors	**First-pass acceptance**	**Numeric/source fidelity**
Prompt + correction cycles	**Total author/QC effort**	**Complexity & exception class**

ACTION → Sample representative tasks; avoid masking failures in pooled averages.

Sources: [FDAAI25] [NIST24]

Governance must convert user feedback into better products

USER FEEDBACK

Document failed or burdensome task(s)

→

STATISTICIAN TRIAGE

Classify impact, identify root cause where possible, bypass is valid

→

DEVELOP + RETEST

Update instructions, examples, controls

→

RELEASE DECISION

Accept, restrict, monitor, or stop

TRACK WHAT MATTERS

Critical errors · Source reconciliation · Total reviewer effort · Correction cycles · Change history

ACTION → Assign use-case owners and require evidence before expanding deployment.

Sources: [FDAEMA26] [NIST24]

THE HUMAN IMPERATIVE

The future of AI in biostatistics must be statistician-led.

01	**START WHERE REVIEW IS TRACTABLE**	Bounded TFL-to-text and well-specified ADaM patterns
02	**DESIGN FOR TRACEABILITY**	Source provenance, numeric checks, controlled approval
03	**EMBED STATISTICAL AUTHORITY**	Requirements, evaluation, error decisions, boundaries

It is critical that the process is statistician led - not only statistician reviewed

Q&A

Primary research and regulatory sources

[NZ23] Noy & Zhang, Science (2023)

https://doi.org/10.1126/science.adh2586

[MARK25] Markey et al., Clinical Trials (2025)

https://doi.org/10.1177/17407745251320806

[METR25] Becker et al., METR developer RCT (2025)

https://metr.org/Early_2025_AI_Experienced_OS_Devs_Study-paper.pdf

[METR26] METR developer productivity update (2026)

https://metr.org/blog/2026-02-24-uplift-update/

[SHIN26] Shin et al., Clinical Pharmacology & Therapeutics (2026)

https://doi.org/10.1002/cpt.70096

[FDAAI25] FDA draft AI guidance (Jan 2025)

https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological

[FDAEMA26] FDA–EMA Good AI Practice principles (Jan 2026)

https://www.fda.gov/about-fda/artificial-intelligence-drug-development/guiding-principles-good-ai-practice-drug-development

[E6R3] FDA ICH E6(R3) GCP (Sep 2025)

https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e6r3-good-clinical-practice-gcp

Standards, traceability, and risk-management sources

[E9R1] FDA ICH E9(R1) estimands (May 2021)

https://www.fda.gov/regulatory-information/search-fda-guidance-documents/e9r1-statistical-principles-clinical-trials-addendum-estimands-and-sensitivity-analysis-clinical

[M11] FDA ICH M11 protocol guidance (May 2026)

https://www.fda.gov/regulatory-information/search-fda-guidance-documents/m11-clinical-electronic-structured-harmonised-protocol

[ADAM] CDISC ADaM standards

https://www.cdisc.org/standards/foundational/adam/primer

[ADAMCONF] CDISC ADaM Conformance Rules v5.0 (2023)

https://www.cdisc.org/standards/foundational/adam/adam-conformance-rules-v5-0

[POPPK] CDISC ADaM popPK Implementation Guide v1.0

https://www.cdisc.org/adam-poppk-implementation-guide-v1-0

[FDASTD26] FDA Study Data Technical Conformance Guide (Oct 2026)

https://www.fda.gov/regulatory-information/search-fda-guidance-documents/study-data-technical-conformance-guide-technical-specifications-document

[NIST24] NIST Generative AI Profile (2024)

https://doi.org/10.6028/NIST.AI.600-1

[NIHCLIN] NIH requirements for trial registration and results reporting

https://grants.nih.gov/policy-and-compliance/policy-topics/clinical-trials/reporting