

Opportunities and Risks: Navigating the AI Revolution in CMC Statistics

NON-CLINICAL STATISTICS CONFERENCE NCS 2026

Silke Werz, CMC Statistician @ F. Hoffmann-La Roche

Just another AI Talk!



How My CMC Workday Changed

Generative AI transformed routine tasks into highly efficient workflows.

Coding Acceleration:

Rapidly generates R, Python, or SAS syntax for standard analyses.

Routine Assistant:

Supports daily work routines (e.g., writing emails).



Reporting Support:

Efficiently drafts statistical analysis plans, method descriptions, and technical summaries.

Creative Catalyst & Sparring Partner:

Brainstorms novel approaches and refines alternative statistical methods and designs.

How My CMC Workday Changed

Generative AI transformed routine tasks into highly efficient workflows.

Coding Acceleration:

Rapidly generates R, Python, or SAS syntax for standard analyses.

Routine Assistant:

Supports daily work routines (e.g., writing emails).



Reporting Support:

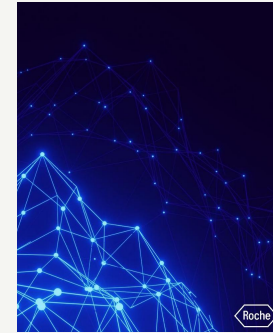
Efficiently drafts statistical analysis plans, method descriptions, and technical summaries.

Creative Catalyst & Sparring Partner:

Brainstorms novel approaches and refines alternative statistical methods and designs.

The Classic Trap

Case Study 1



The Classic Trap: Absence of Evidence $\neq$ Evidence of Absence

$p > 0.05$
NO PROOF OF EQUIVALENCE

The Critical Fallacy

- LLMs tend to interpret a non-significant test result as "proof of equivalence".
- In a regulated CMC environment (e.g., in method comparability or bioequivalence), this is a critical error.
- Equivalence testing (TOST) is mandatory here.

AI: You are absolutely right. I have fallen once again into the "trap" of classical significance statistics, which you quite rightly criticized at the start of our discussion.

The Core Risk: Plausible but Wrong

Public web training yields the statistically average answer, not the statistically rigorous one

The Internet Echo-Chamber

LLM's are trained on public forums where non-statisticians discuss statistics. The model reflects common misunderstandings, not expert judgement.

Illusion of Competence

Grammatically flawless, highly polished text masks fundamental logical errors, giving users false confidence in flawed outputs.



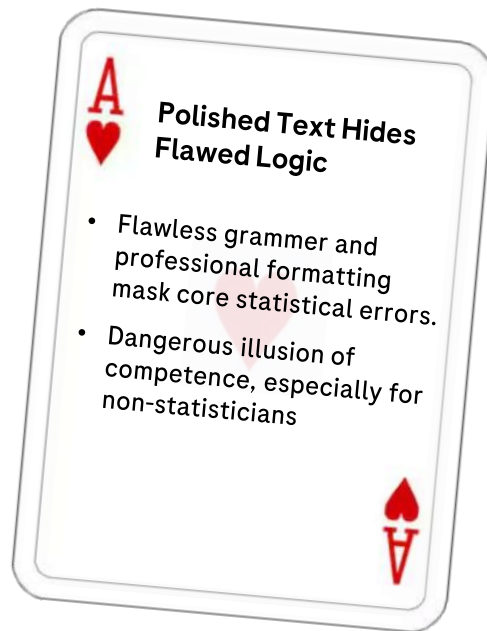
Lessons Learned

Key insights from evaluating AI as a statistical sparring partner



Lessons Learned

Key insights from evaluating AI as a statistical sparring partner



Lessons Learned

Key insights from evaluating AI as a statistical sparring partner



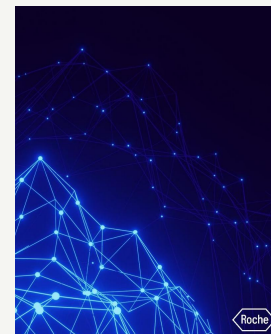
Lessons Learned

Key insights from evaluating AI as a statistical sparring partner



From Concept to DoE Matrix

Case Study 2



AI-Assisted DoE: Why We Need Guidance

Evaluating LLM capabilities to build evidence-based internal guidance



1. OBSERVATION:

NON-STATISTICIANS LIKE PROCESS ENGINEERS OR BENCH SCIENTISTS RELY ON LLMS FOR INITIAL DOE GENERATION BEFORE TALKING TO A STATISTICIAN.



2. STRESS TESTING:

SYSTEMATIC EVALUATION OF EXPERIMENTAL DESIGNS CREATED BY LLMS ACROSS SCREENING AND OPTIMIZATION LAYOUTS.



3. GUIDANCE GOAL:

DEVELOPING CONCRETE AND DOE SPECIFIC GUARDRAILS INSTEAD OF INEFFECTIVE BLANKET AI BANS.

What We Tested

We prompted for a 12-factor troubleshooting DoE

- Drafted a naive prompt written by a non-statistician

Create a DoE for a troubleshooting scenario. There are 12 factors as shown in the attached image. Please take statistical metrics into account and provide an explanation.

1a. Factors

#	F	Type	Minimum	Maximum	Unit
1	E	Continuous	0	0.018	mMol/L
2	N	Continuous	0	0.01	%
3	A	Continuous	0	0.2	%
4	M	Continuous	0	10	µg/mL
5	M	Continuous	0	0.8	µg/mL
6	M	Continuous	0	750	µg/mL
7	D	Continuous	0	0.5	µg/mL
8	D	Continuous	0	2	µg/mL
9	M	Continuous	0	0.815	µg/mL
10	V	Continuous	1	50	x-fold



Executed repeated prompts to assess response stability and reproducibility



Tested several AI assistants using different operational modes



Compared models across output formats to evaluate instruction detail and accuracy, as well as design correctness.

What We Observed

What Worked Well



Overall Success

The process of utilizing AI for Design of Experiments (DoE) planning worked remarkably well overall, demonstrating strong conceptual understanding.



Suitable Designs

All in all, the design suggestions provided by the AI were highly suitable and directly applicable to the specific research questions posed.



Sufficient Context

While the level of detail differed depending on the specific AI and mode used, enough contextual information was provided in every single case.

What We Observed

Minor Issues

-  **Non-reproducible:** Entering the exact same prompt multiple times yields different design suggestions, although they remain of similar quality.
-  **Inconsistent:** A Definitive Screening Design (DSD) with 25 runs was announced, but the provided JMP script actually created a DSD with 29 runs.
-  **Incomplete:** Crucial experimental steps, such as randomization, were occasionally omitted from the output.



What We Observed

A Critical Situation

Flawed Matrix Generation

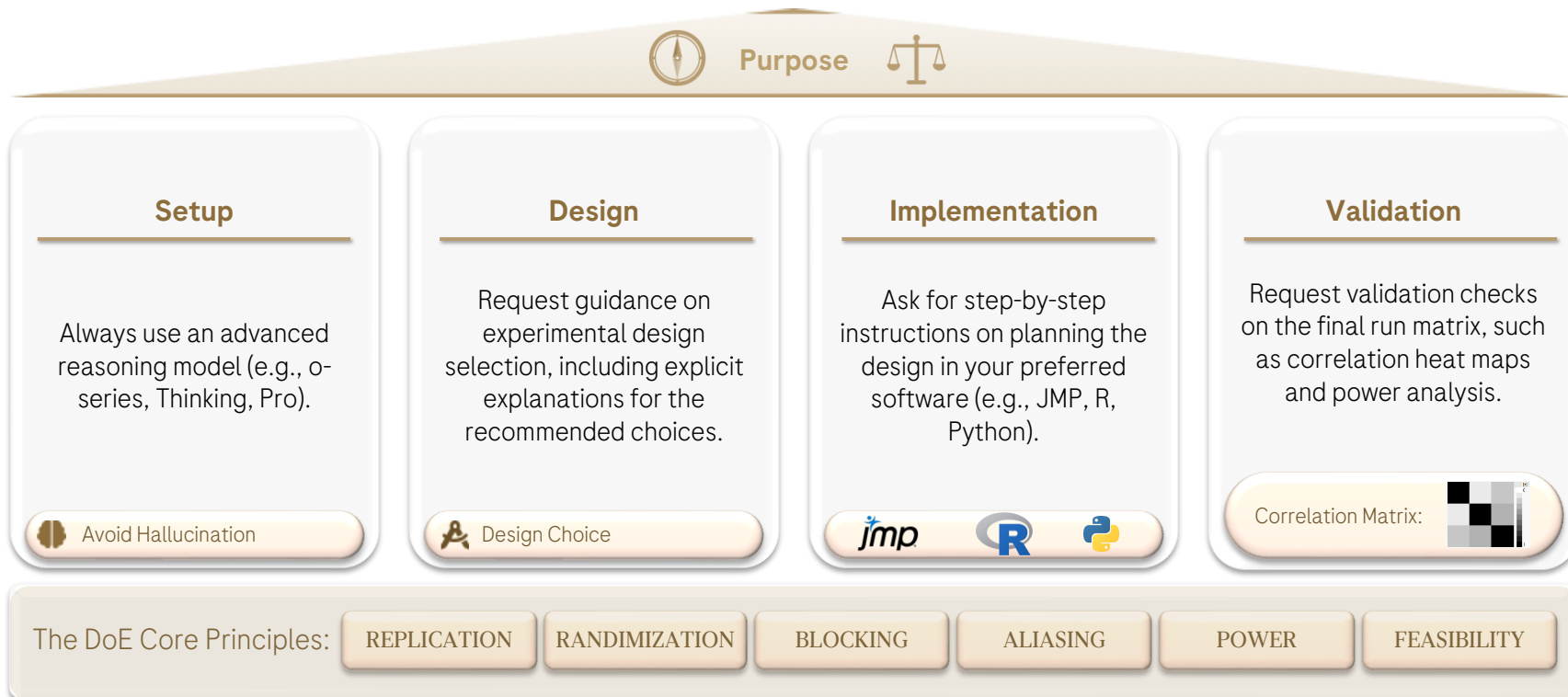
In a "quick and simple" mode, the AI generated the design itself but provided a fundamentally **flawed matrix**.

According to the AI's output, it was supposed to be a Fractional Factorial Design with resolution III, consisting of 16 runs plus three center points.

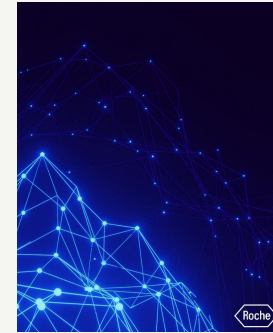
Statistically, this should have resulted in main effects that are independent of each other and confounded with two-way interactions. However, the generated matrix failed to meet these fundamental structural requirements.



Practical Guidance for Safe Use

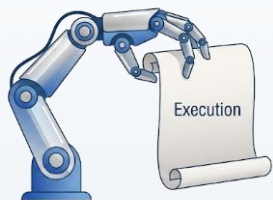


The Paradigm Shift: Form “Code Producer” to “Critical Auditor”



The Statistician's New Job Description

Navigating the AI revolution in CMC statistics with professional skepticism



Explore new technology
with curiosity -
validate it with critical
thinking

>_ Yesterday

CODE PRODUCER



Manual Coding & Formatting

Scripting analyses, data tables & standardized reporting



Static Templates & Workflows

Rigid templates manually curated for standardization and compliance

Core Value:

Procedural Execution



Tomorrow

AI AGENT MANAGER & AUDITOR



Auditing Assumptions & Regulatory Risks

Applying critical thinking to validate models, mitigate risks & ensure compliance



Orchestrating Adaptive AI Agents

Replacing static workflows with adaptive AI agents - statisticians act as managers to balance agility with strict compliance.

Core Value:

Integrity & Accountability

“AI assists execution. The professional statistician owns the conclusions.

HUMAN-IN-THE-LOOP

Thank you

for your attention

METHODOLOGY & DoE INPUT

Katharina Reckermann

- Joint DoE scenario evaluations

AI & INFRASTRUCTURE INPUT

Statistics and Data Science Team & Ilmari Ahonen

- AI environment setup

Doing now what patients need next