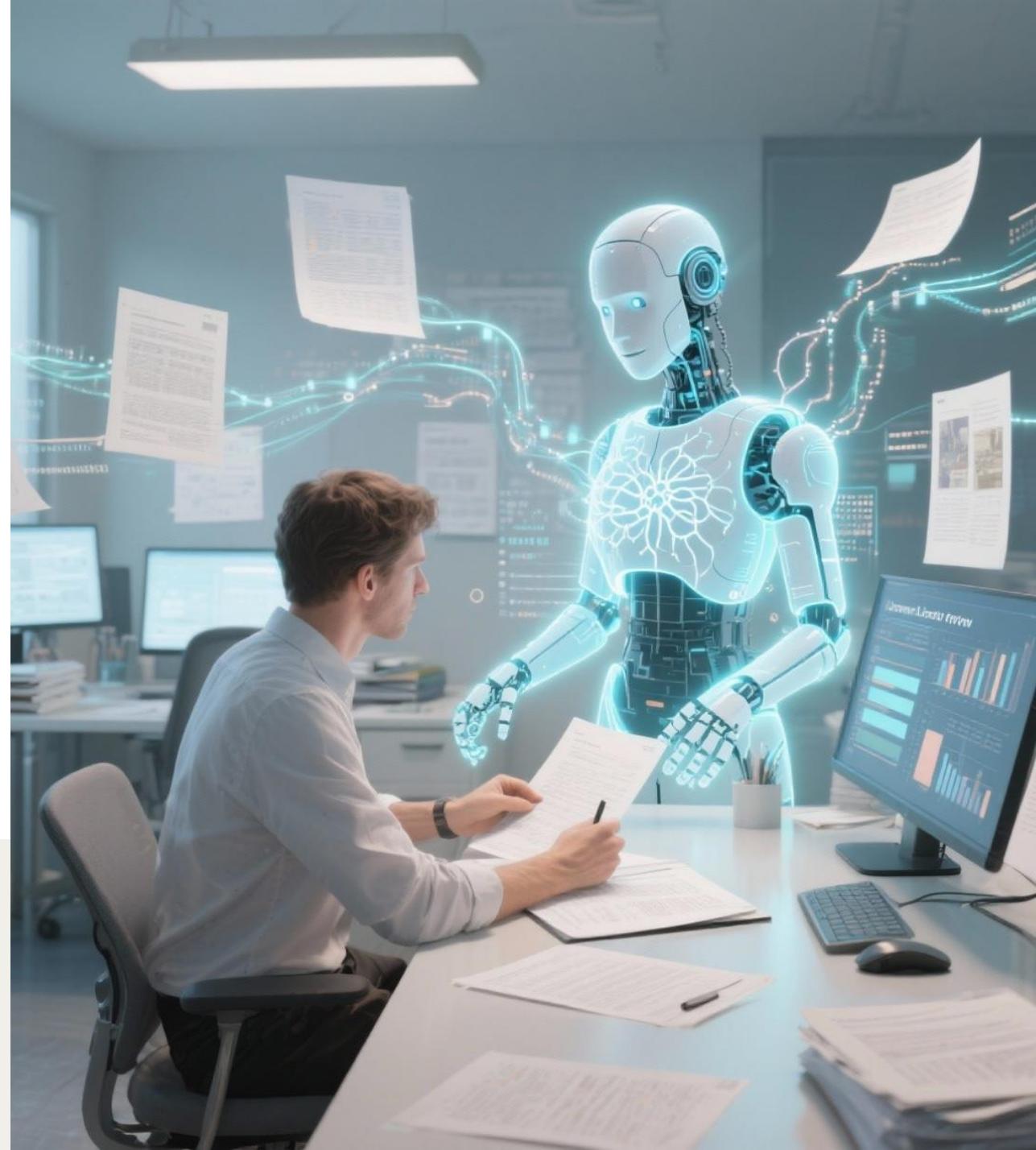


GenAI in SLRs: From Dual Review to Human-AI Collaboration

November 11, 2025

ISPOR Glasgow

Sven Klijn





The opinions expressed in this presentation and on the following slides are solely those of the presenter and do not necessarily represent opinions of Bristol Myers Squibb.

Dual human review

Human 1

Formulating the research question



PICO definition



Search query



Title/abstract screening



Full text screening



Data extraction



Reporting

Human 2

Title/abstract screening



Full text screening



Data extraction

Human 3

Arbitration

Arbitration

Arbitration

The gold standard is not perfect



Time-
consuming



Limited
scalability



Screening
fatigue

“The total error rate
(false inclusion and false exclusion)
was 10.76% (95% CI: 7.43% to 14.09%)”

Wang et al. (2020). Error rates of human reviewers during abstract screening in systematic reviews. DOI: 10.1371/journal.pone.0227742

Human-AI Collaboration

Human 1

Formulating the research question



PICO definition



Search query



Title/abstract screening



Full text screening



Data extraction



Reporting

AI

Title/abstract screening



Full text screening



Data extraction

Human 2

Arbitration

Arbitration

Arbitration



AI Governance



Model Documentation

- Architecture
- Training data
- Performance metrics
- Logging



Life Cycle Maintenance

- Version control
- Bias & fairness assessments:
Systematic evaluation of outputs
- Reproducibility

Best practice principle 6:

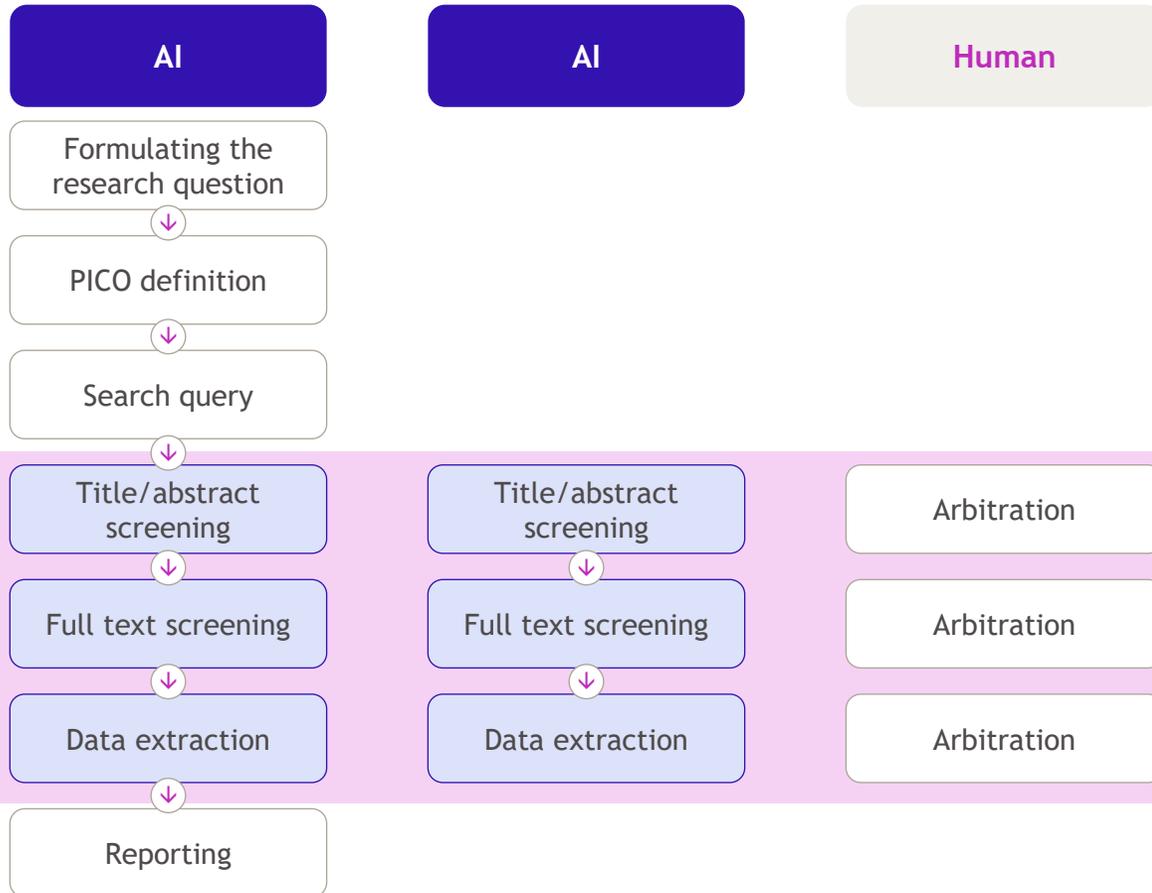
“Ensure GenAI-assisted analyses are reproducible and auditable where possible.”

NICE (2025). Generative Artificial Intelligence (GenAI) best practice principles

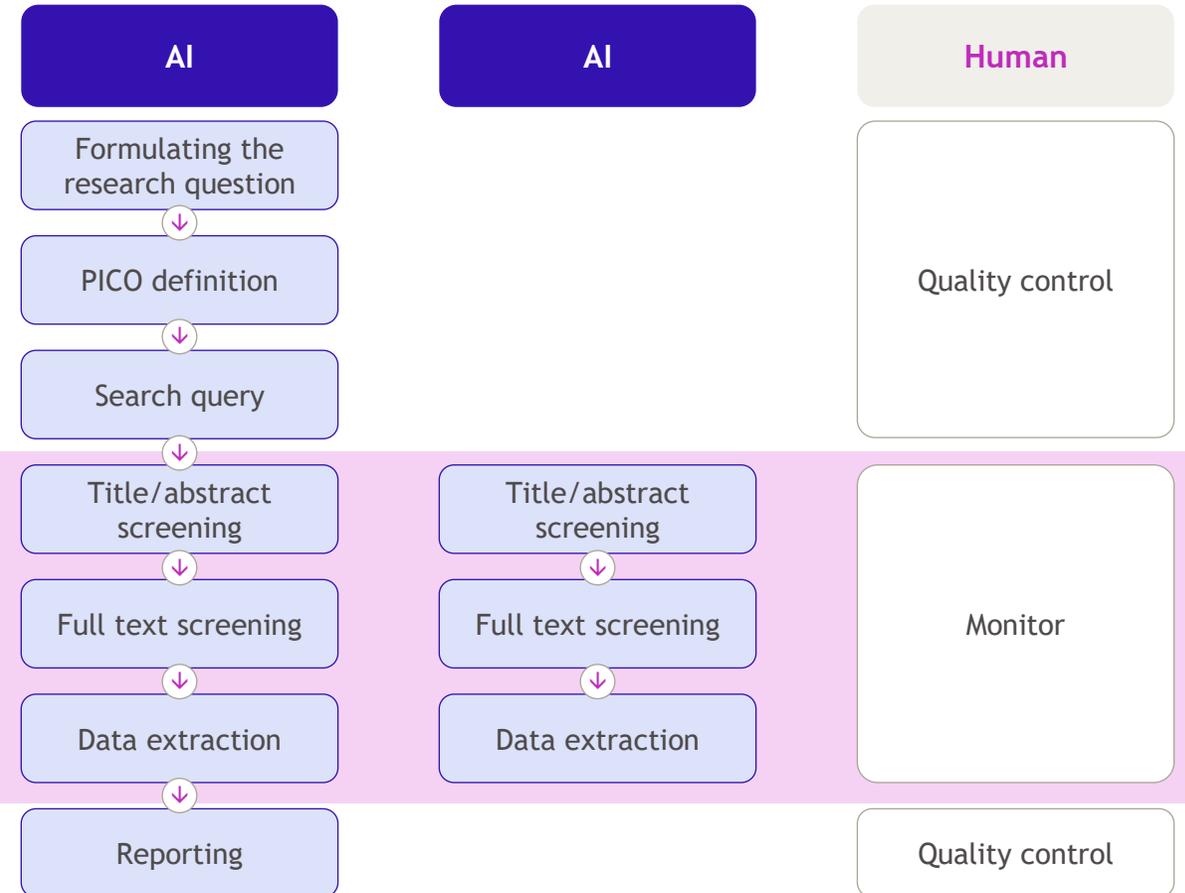


Alternative models

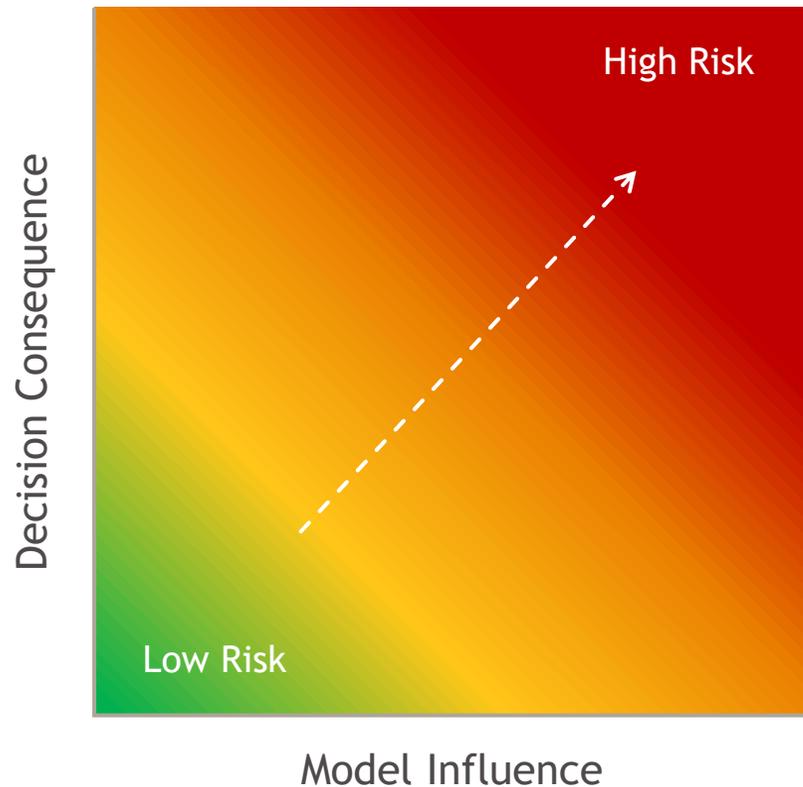
Human Arbitration



Human Monitoring



Model risk matrix



Decision consequence depends on:

- Type of publication
- Type of error made:
false inclusion vs. false exclusion



Model influence depends on:

- Model used:
Human-AI / Human Arbitration / Human Monitoring
- Guardrails and mitigation strategies



Reflection



Discussion required on
where the bar should be set



Strong model governance is critical



Role of the human is essential,
but differs by implementation model