ASA-BIOP
Safety Scientific Working Group
3Q Webinar

AI Agents in Quantitative Safety Evaluation

Emerging Applications in Pharmacovigilance and Causal Inference

Wei Wang*, William Wang

Biostatistics and Research Decision Sciences (BARDS)

Merck & Co., Inc., Rahway, NJ, USA

September 09, 2025

On Behalf of the ASA-BIOP Safety Scientific Working Group

Work Stream 2 (Statistical Methodology)

Disclaimer

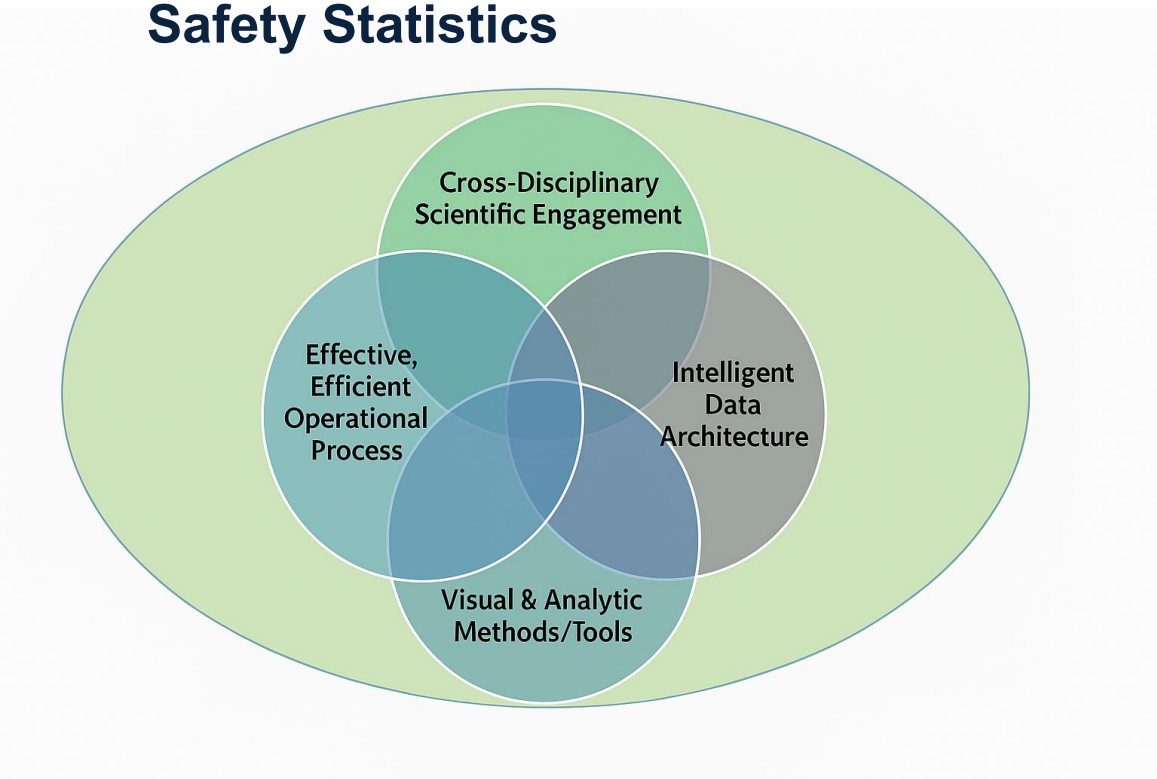
The content of this presentation and related discourse during this ASA safety SWG webinar is not necessarily reflective of the positions, policies, or practices of the presenter and working group members' employers.

Motivation

Quantitative Safety Evaluation – What and Why?

- **What is it?**
  - ❑ A systematic approach using statistical and computational methods to assess drug safety throughout the entire product life cycle – from early phase to late phase and post-marketing surveillance (Jiang & Xia, 2014)
  - ❑ A cross-disciplinary approach that involves collaboration among statisticians, safety scientists, clinicians, and regulators (Wang et al., 2022)
  - ❑ The goal is to objectively evaluate the safety profile, detect and quantify safety signals, and support benefit-risk assessment (Wang et al., 2022)
- **Why is it important?**
  - ❑ Regulatory agencies require robust safety evidence for approval and post-marketing surveillance (EMA, 1997)
  - ❑ Support proactive risk management strategies (EMA, 2024)

Challenges and Opportunities



- **Data Complexity:**
  - ❑ Integration of diverse sources: clinical trials, claims, spontaneous reports, EHRs, registries
  - ❑ Variability in formats, standards, and data quality
- **Statistical Limitations:**
  - ❑ Premarketing trials often underpowered for rare or long-term adverse events
  - ❑ Lack of gold-standard methods for causality assessment in real-world data
- **Operational Pressure:**
  - ❑ Accelerated timelines for regulatory submissions
  - ❑ Rising expectations for transparency, reproducibility, and stakeholder communication
- **Collaborative Innovation:**
  - ❑ Need for integrated expertise across biostatistics, epidemiology, clinical science, informatics, and regulatory affairs

Wang, W., Munsaka, M., Buchanan, J., & Li, J. (Eds.). (2022). Quantitative drug safety and benefit-risk evaluation:
Practical and cross-disciplinary approaches. Chapman & Hall/CRC

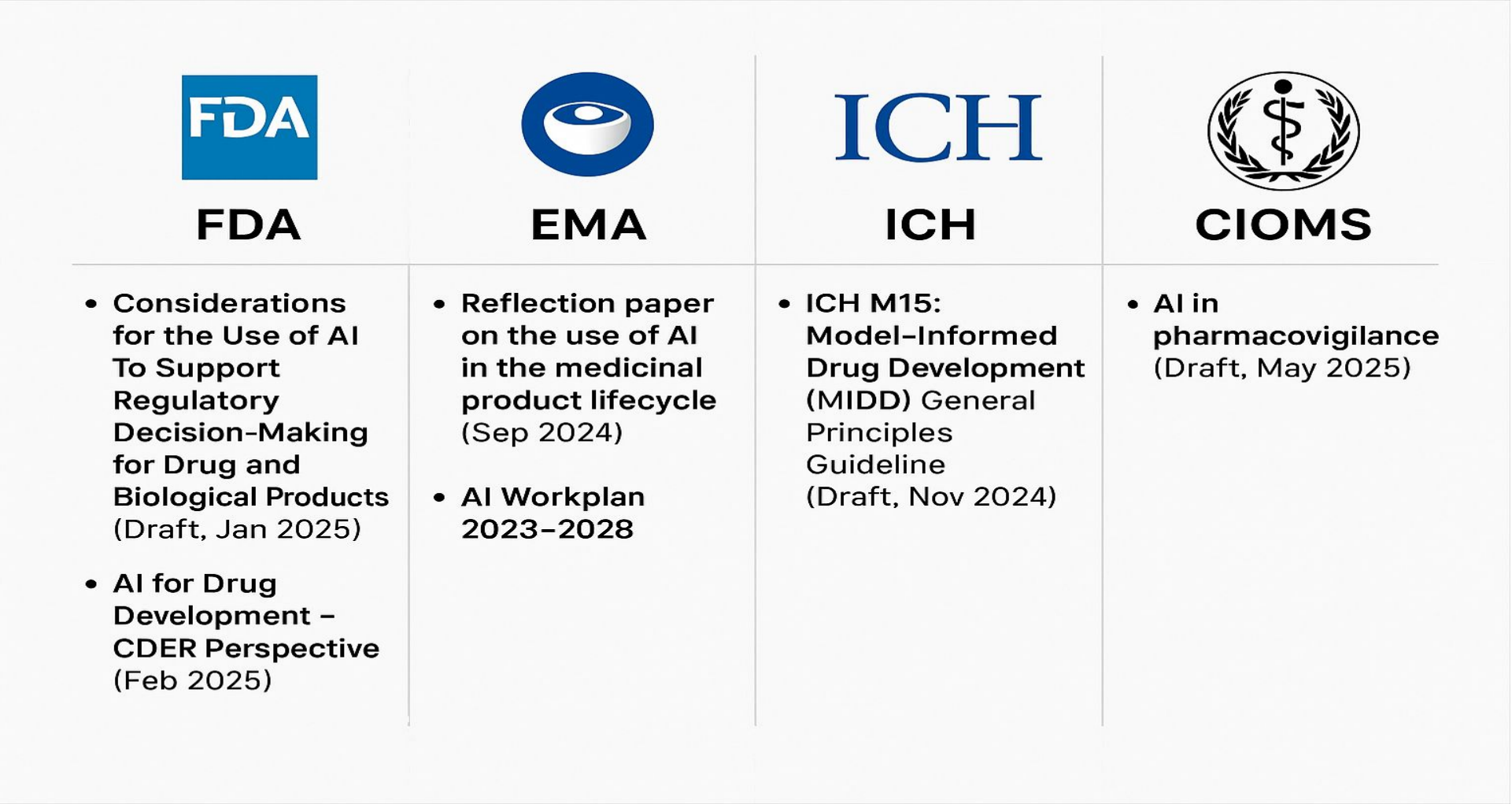
Why Can AI Agents Help?

- **AI Agent:**
  - An intelligent *message transformer* that can respond to messages through interaction with Large Language Models (LLMs), and may also be equipped with *tools* and *external documents/data*
- **Advantages of AI Agent/LLM:**
  - Capability in text/visual understanding and generation
  - Interdisciplinary expertise
- **Disadvantages of AI Agent/LLM:**
  - "Built-in" knowledge is based on (mostly public) data sources it has "seen" during pre-training
  - Being a probabilistic next-token predictor, it may produce incorrect or unreliable results – hallucination

Key Areas Where AI Agents May be Helpful

- **Data Integration and Pre-Processing:**
  - ❑ Aggregate data from different sources (e.g., clinical trials, EHR, literature, etc.) and normalize heterogeneous formats and apply standard vocabularies (e.g., MedDRA, WHO-DD)
  - ❑ Automate data cleaning and quality checks
- **Signal Detection and Monitoring:**
  - ❑ Continuously scan large safety databases for detecting emerging safety signals earlier than traditional methods (e.g., disproportionality analysis)
  - ❑ Enable real-time pharmacovigilance by leveraging external resources (e.g., drug labels, medical literature, openFDA)
- **Causal Inference and Risk Management**
  - ❑ Assist in study design (e.g., target trial emulation, confounder identification) and suggest appropriate statistical methods

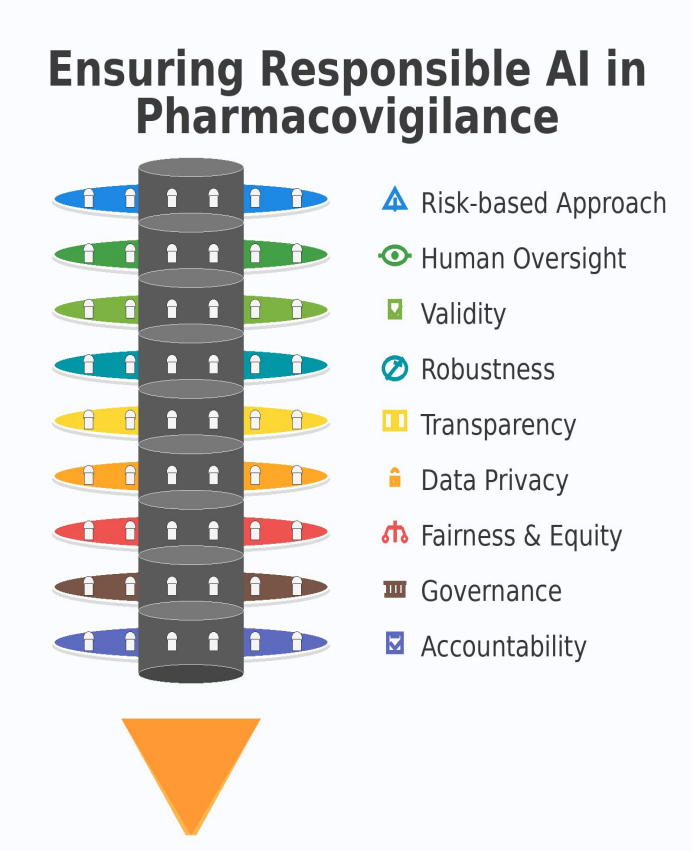
Regulatory Guidance on AI Applications in Drug Development



Generated by Microsoft 365 Copilot

CIOMS Working Group Report on AI in Pharmacovigilance (May 2025)

CIOMS Working Group XIV has released a draft report on AI in Pharmacovigilance, which outlines the key principles, applications, and future directions



Generated by Microsoft 365 Copilot

CIOMS-WG-XIV_Draft-report-for-Public-Consultation_1May2025.pdf

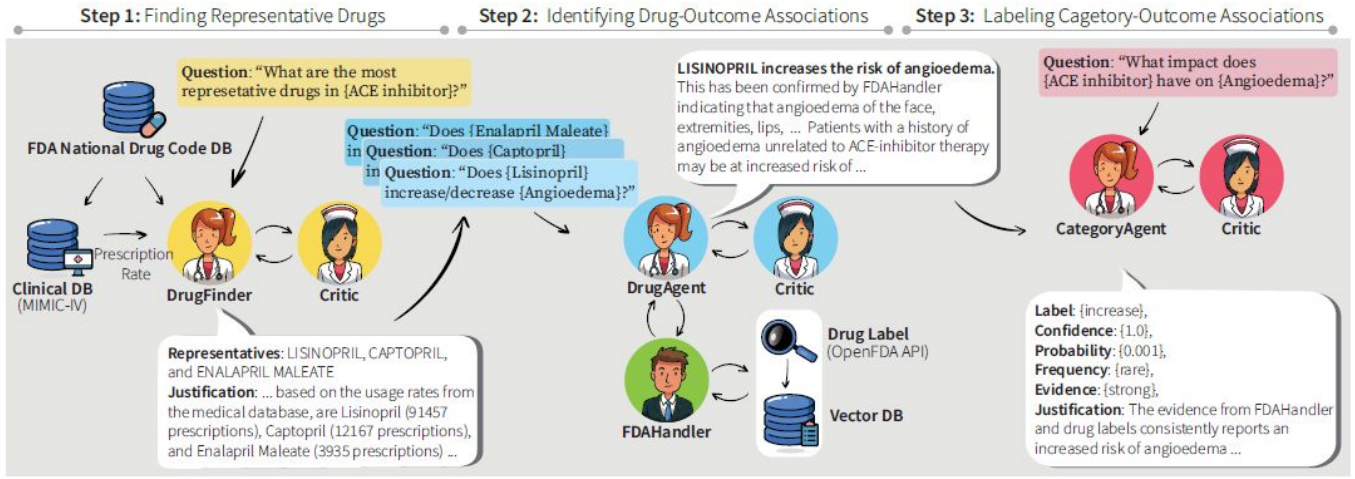
Case #1: MALADE in Pharmacovigilance

MALADE=Multiple **A**gents powered by **L**LMs for **AD**E **E**xtraction

MALADE: Orchestration of LLM-powered Agents with Retrieval Augmented Generation for Pharmacovigilance

Jihye Choi*[1], Nils Palumbo*[1], Prasad Chalasani[2], Matthew M. Engelhard[3], Somesh Jha[1,2], Anivarya Kumar[3], David Page[3]

[1]University of Wisconsin-Madison, [2]Langroid, [3]Duke University

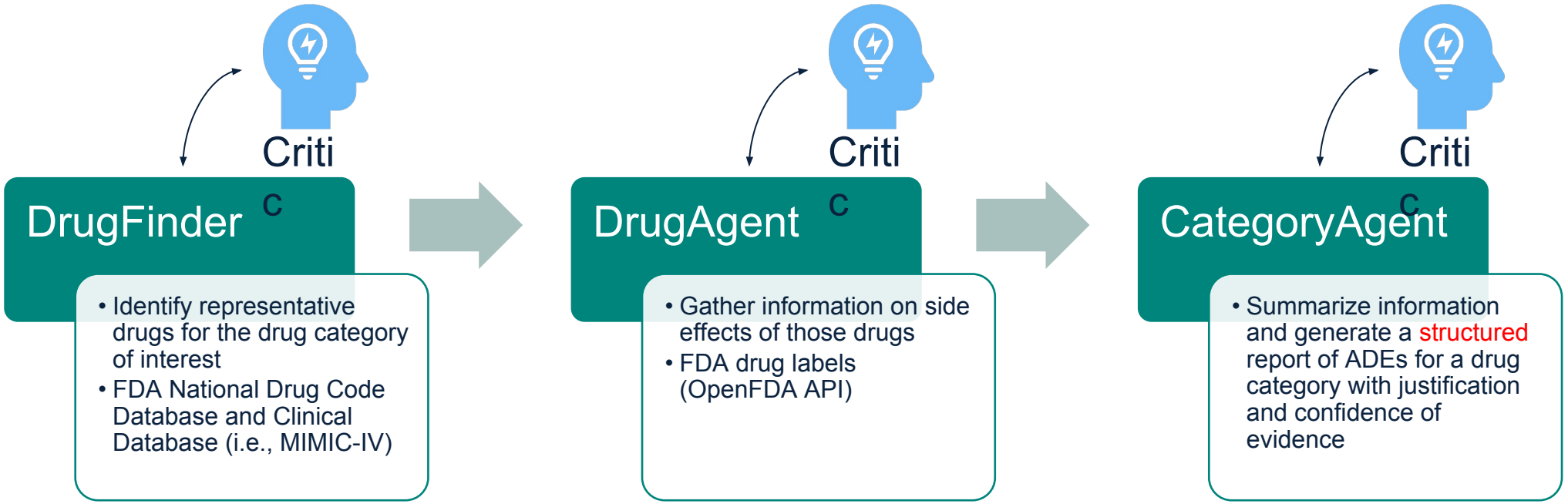


- **Challenges:**
  - ❑ Variability in names of drugs and outcomes
  - ❑ Information is not always in a reader-friendly format and oftentimes is buried in large amounts of narrative texts
- **Other Attempts:**
  - ❑ Traditional Natural Language Processing (NLP)/Deep Learning methods are not good at text understanding and generation
  - ❑ Off-the-shelf agents (e.g., chatGPT) may not be trained with required domain knowledge

Task: *identify adverse drug events (ADEs) from drug labels for a drug category*

[2408.01869] MALADE: Orchestration of LLM-powered Agents with Retrieval Augmented Generation for Pharmacovigilance

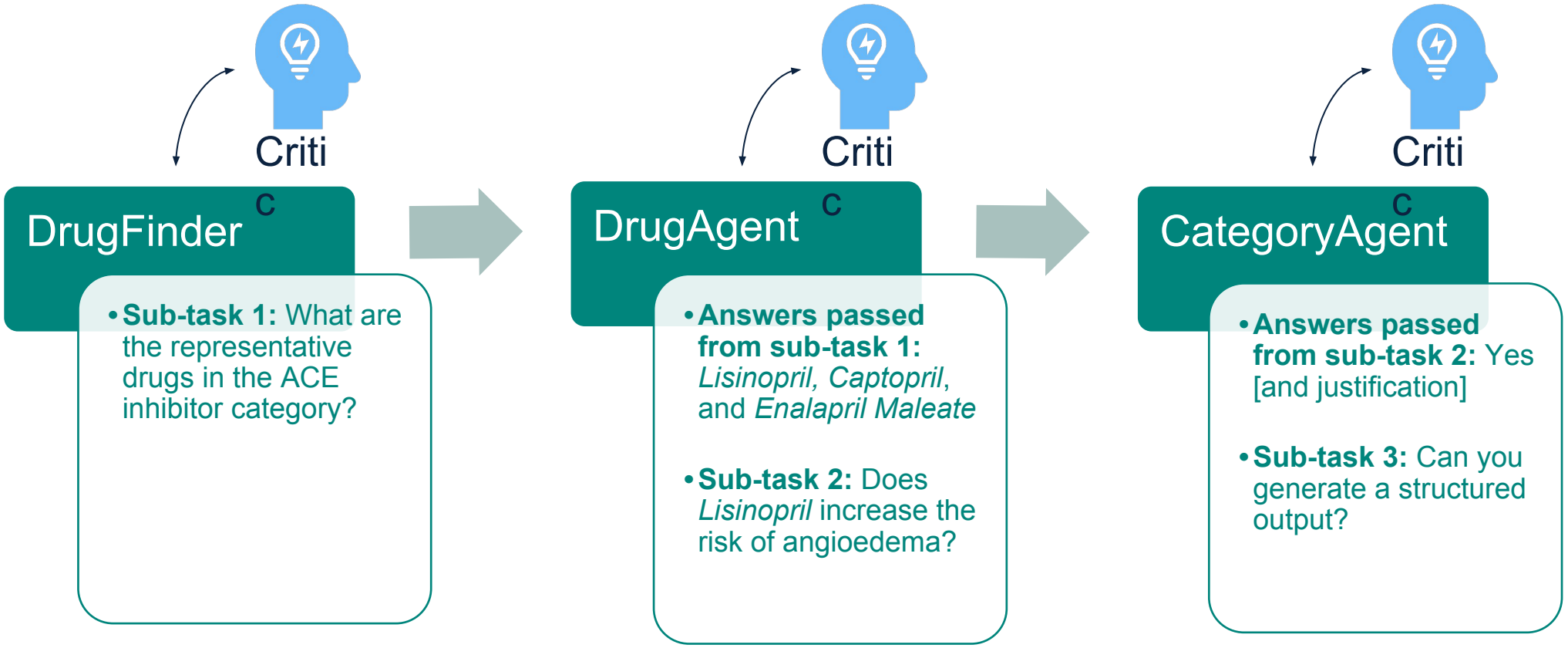
How MALADE Works



- **MALADE unique features:**
  - ❑ A multi-agent system, in which each of them works on a sub-task
  - ❑ Instructed to produce structured outputs/responses, which are called *tools*, typically in JSON format with various pre-specified fields, such as code, SQL query, parameters of an API call, etc.
  - ❑ Equipped with Retrieval Augmented Generation (RAG), agents fetch relevant external documents and use them to generate context-grounded responses, reducing knowledge gaps

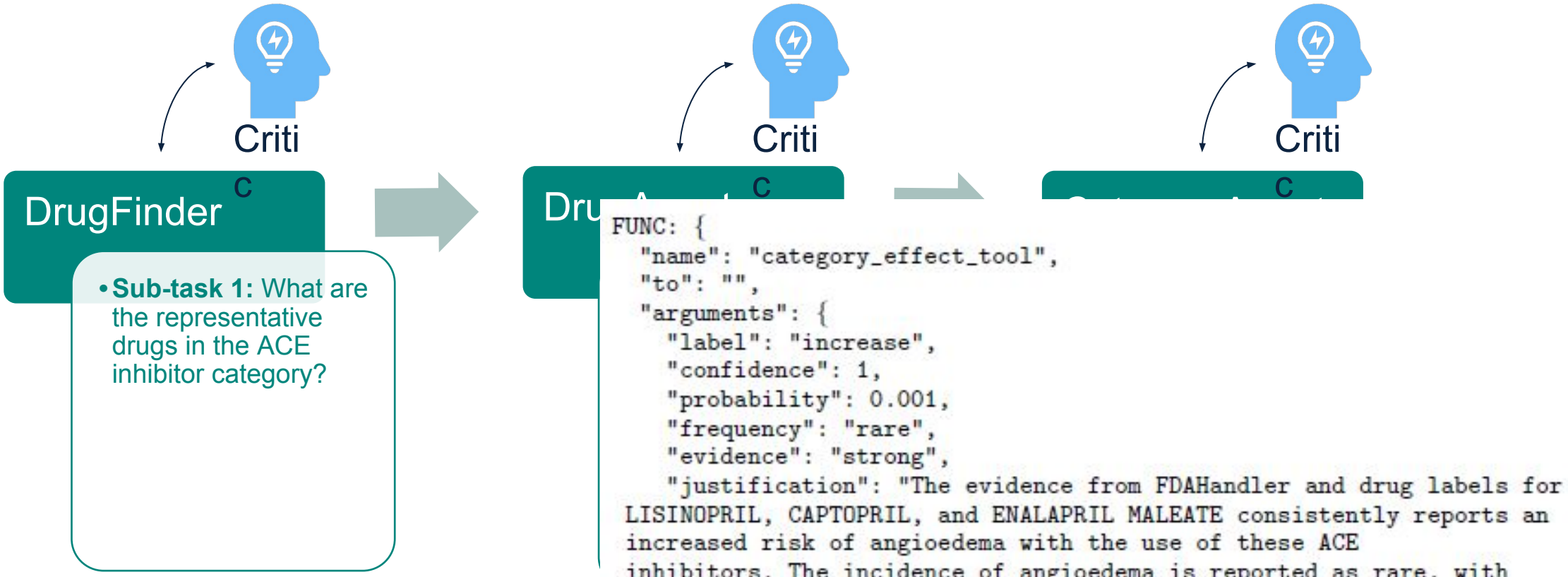
DEMO

Task: Does the angiotensin converting enzyme (ACE) inhibitor increase the risk of angioedema?



DEMO

Task: Does the angiotensin converting enzyme (ACE) inhibitor increase the risk of angioedema?



Case #2: Causal Copilot in Real-World Studies

Large Language Models as Co-Pilots for Causal Inference in Medical Studies

Ahmed Alaa
UC Berkeley and UCSF

Rachael V. Phillips
UC Berkeley

Emre Kıcıman
Microsoft Research

Laura B. Balzer
UC Berkeley

Mark van der Laan
UC Berkeley

Maya Petersen
UC Berkeley

Despite growing recognition of real-world data (RWD) as a valuable resource, establishing causal effects from RWD remains a significant challenge.

RWD Challenges

- Selection bias
- Intercurrent events
- Informative missingness
- Treatment by indication
- High dimensional covariates
- Outcome measurement error
- Statistical model misspecification
- Differences between external controls and single-arm RCT

Causal frameworks support RWD causal inference and reliable RWE generation

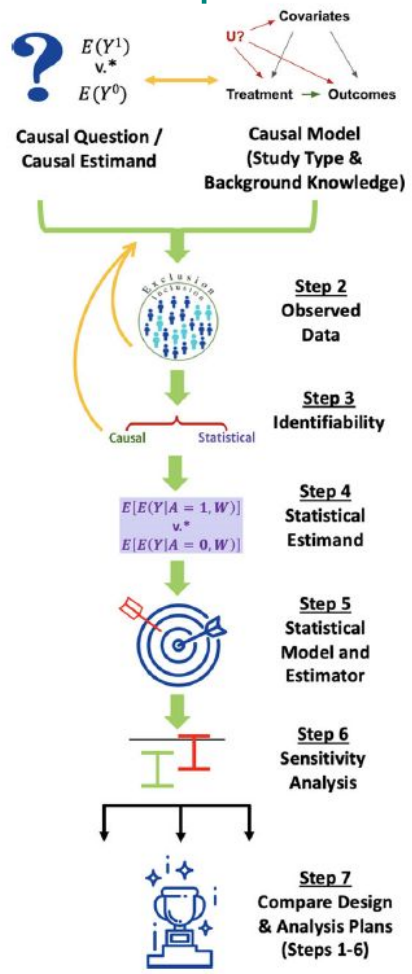
The *Roadmap* — A Unified Causal Framework

- ★ Step-by-step guidance to prespecify study design and analysis plan that is generally applicable to any use case along the study spectrum
- ★ Builds on decades of research on careful study design, and compatible with existing causal frameworks
- ★ Yields high-quality estimates of causal effects (e.g., using RWD) when possible; otherwise, provides insight into future refinements to improve credibility
- ★ Flexible estimation and dimension reduction with machine learning, and formal statistical inference
- ★ Model-free sensitivity analysis

[2407.19118] Large Language Models as Co-Pilots for Causal Inference in Medical Studies

Causal Copilot is Promising but Challenging

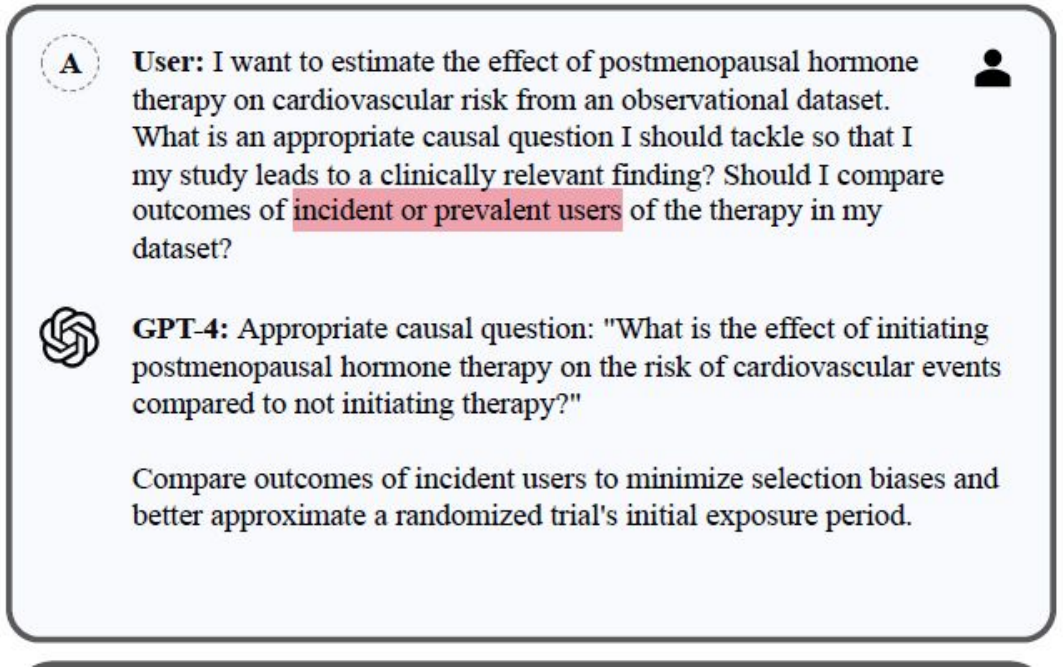
Causal Roadmap



https://doi.org/10.1017/cts.2023.635

- **Causal copilot could assist researchers along the causal roadmap:**
  - ❑ Clarify causal question and estimand
  - ❑ Suggest data sources and structure
  - ❑ Check identifiability assumptions
  - ❑ Recommend statistical estimands
  - ❑ Suggest models and diagnostics
  - ❑ Propose sensitivity analyses
- **Research direction and challenges in developing a causal copilot:**
  - ❑ Ground the copilot on domain knowledge (e.g., well established causal framework, regulatory guidance), rather than relying on off-the-shelf LLMs
  - ❑ Finetune the copilot with “instruction-tuning” data, containing demonstrations of RW studies, researcher instructions, and ideal responses
  - ❑ Keep humans in the loop by utilizing reinforcement learning from human feedback (RLHF)

Example: Causal Copilot Can Clarify the Causal Question



- **Clinical question: does postmenopausal hormone replacement therapy (HRT) have cardiovascular benefits?**
  - ❑ In 1980s and 1990s, several real-world studies suggested positive results, which were later contradicted by multiple RCTs, including the Women's Heath Initiative (WHI) study
  - ❑ The underlying causal question was flawed by comparing incidence of cardiovascular events in women who were currently on HRT and those who were never on HRT (prevalent users), whereas RCT compared HRT initiators and non-initiators (incident users)

- **(Appropriate) Causal question/estimand: What is the effect of initiating postmenopausal HRT on the risk of cardiovascular events compared to not initiating therapy?**

Conclusions

Concluding Remarks

- AI agents are not just tools – they are ‘collaborators’ in advancing drug safety science
- Quantitative safety evaluation benefits from AI’s ability to integrate diverse data, detect signals early, and support causal reasoning
- Regulatory bodies are actively shaping the landscape, signaling a future where AI is embedded in pharmacovigilance and real-world evidence generation
- Case studies like MALADE and Casual Copilot demonstrate both the promise and the challenges of deploying AI responsibly
- Collaboration across disciplines is essential to harness AI’s full potential

Let’s continue to innovate responsibly, keeping patient safety and scientific rigor at the core

References

1. Jiang, Q., & Xia, H. A. (Eds.). (2014). *Quantitative evaluation of safety in drug development: Design, analysis and reporting*. CRC Press.
2. European Medicines Agency (EMA). (2010). Benefit-Risk Methodology Project: Work Package 2 report – Applicability of current tools and processes for regulatory benefit-risk assessment. https://www.ema.europa.eu/en/documents/report/benefit-risk-methodology-project-work-package-2-report-applicability-current-tools-and-processes-regulatory-benefit-risk-assessment_en.pdf
3. EMA (1997). ICH E2C(R1) Clinical Safety Data Management: Periodic Safety Update Reports (PSURs). https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-2-c-r1-clinical-safety-data-management-periodic-safety-update-reports-marketed-drugs-step-5_en.pdf
4. EMA. (2024). Guideline on good pharmacovigilance practices (GVP) Module XVI – Risk minimisation measures (Rev 3). https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/guideline-good-pharmacovigilance-practices-gvp-module-xvi-risk-minimisation-measures-rev-3_en.pdf
5. Wang, W., Munsaka, M., Buchanan, J., & Li, J. (Eds.). (2022). Quantitative drug safety and benefit-risk evaluation: Practical and cross-disciplinary approaches. Chapman & Hall/CRC. https://www.routledge.com/Quantitative-Drug-Safety-and-Benefit-Risk-Evaluation-Practical-and-Cross-Disciplinary-Approaches/Wang-Munsaka-Buchanan-Li/p/book/9781032191119
6. U.S. Food and Drug Administration (FDA). (2025, January). *Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products* (Draft Guidance). https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological
7. U.S. FDA. (2025, January). *Artificial intelligence-enabled device software functions: Lifecycle management and marketing submission recommendations* (Draft Guidance). https://www.fda.gov/regulatory-information/search-fda-guidance-documents/artificial-intelligence-enabled-device-software-functions-lifecycle-management-and-marketing
8. U.S. FDA, Center for Drug Evaluation and Research. (2025, February). Artificial intelligence for drug development – CDER perspective. https://www.fda.gov/about-fda/center-drug-evaluation-and-research-cder/artificial-intelligence-drug-development
9. EMA. (2024, September). Reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle. https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle_en.pdf
10. EMA & Heads of Medicines Agencies. (2023, November). *Multi-annual artificial intelligence workplan 2023–2028*. https://www.ema.europa.eu/en/documents/work-programme/multi-annual-artificial-intelligence-workplan-2023-2028-hma-ema-joint-big-data-steering-group_en.pdf
11. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. (2024, November). M15: General principles for

References

13. Choi, J., Palumbo, N., Chalasani, P., Engelhard, M. M., Jha, S., Kumar, A., & Page, D. (2024). MALADE: Orchestration of LLM-powered Agents with Retrieval Augmented Generation for Pharmacovigilance. arXiv. https://doi.org/10.48550/arXiv.2408.01869
14. Alaa, A., Phillips, R. V., Kıcıman, E., Balzer, L. B., van der Laan, M., & Petersen, M. (2024). Large Language Models as Co-Pilots for Causal Inference in Medical Studies. arXiv. https://doi.org/10.48550/arXiv.2407.19118
15. Dang, L. E., Gruber, S., Lee, H., Dahabreh, I. J., Stuart, E. A., Williamson, B. D., Wyss, R., Díaz, I., Ghosh, D., Kıcıman, E., Alemayehu, D., Hoffman, K. L., Vossen, C. Y., Huml, R. A., Ravn, H., Kvist, K., Pratley, R., Shih, M.-C., Pennello, G., Martin, D., Waddy, S. P., Barr, C. E., Akacha, M., Buse, J. B., van der Laan, M., & Petersen, M. (2023).A causal roadmap for generating high-quality real-world evidence. Journal of Clinical and Translational Science, 7, e212. https://doi.org/10.1017/cts.2023.635

Thank you & Questions?