Artificial Intelligence & Causal Inference in Environmental Health Studies

Marianthi-Anna Kioumourtzoglou, ScD
mk3961@cumc.columbia.edu

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH ENVIRONMENTAL HEALTH SCIENCES

HEI Meeting
Austin, TX
05.04.2025

Causal Inference

- No epi study aims to uncover “associations;” the interest is in understanding the causal effect of an exposure on an outcome
- However, due to various biases, causal inference can be challenging

Causal Inference

- No epi study aims to uncover “associations;” the interest is in understanding the causal effect of an exposure on an outcome
- However, due to various biases, causal inference can be challenging

- Causal inference methods can help:
  - Evaluate explicitly counterfactual scenarios
  - Provide robustness against model misspecification
  - Yield effect estimates that can be interpreted causally under appropriate assumptions

Machine Learning & Causal Inference

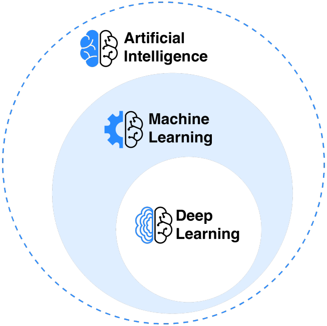
- ML aims to maximize predictive accuracy
- Accommodating high-dimensional, complex structures
- Prediction $\neq$ causal inference

Machine Learning & Causal Inference

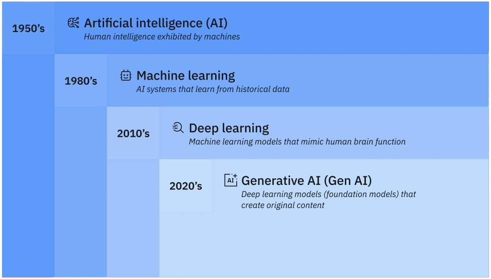
- ML aims to maximize predictive accuracy
- Accommodating high-dimensional, complex structures
- Prediction $\neq$ causal inference

- ML algorithms are being used to increase flexibility in modeling complex relationships while preserving valid inference
- E.g., targeted maximum likelihood estimator (TMLE; van der Laan and Rubin,2006)
  - Integrates ML & statistical inference to reduce bias in efficient & robust estimation

ML ⊂ AI



NASA


1950's
Artificial intelligence (AI)
Human intelligence exhibited by machines
1980's
Machine learning
AI systems that learn from historical data
2010's
Deep learning
Machine learning models that mimic human brain function
2020's
Generative AI (Gen AI)
Deep learning models (foundation models) that create original content

IBM

Deep learning: multi-layer neural networks that can learn features automatically from data

AI & Causal Inference

- Growing interest in deep learning for causal inference, including in methods that
    - Infer causal structure (e.g., using variational autoencoders)
    - Learn balanced or disentangled representations that help reduce confounding or simulate interventions
- Still in nascent stages
- One example approach is GANITE
    - Generative Adversarial Networks for Inference of Treatment Effects (Yoon et al., 2018)
    - Uses GANs to estimate potential outcomes and individual treatment effects

Important Considerations

- These methods are data hungry
    - Although our data are becoming bigger, certain studies (e.g., using biomarkers) still have quite modest sample sizes

- Fair AI (algorithmic fairness)
    - Historical and/or structural biases can be perpetuated
    - Same issue as in ML

Important Considerations (cont'd)

- Interpretable & Explainable AI
    - Deep learning is a black box
    - We want to understand the AI solution and how it got it
    - Could apply post-hoc tools, e.g., SHAP, showing feature contributions[1]

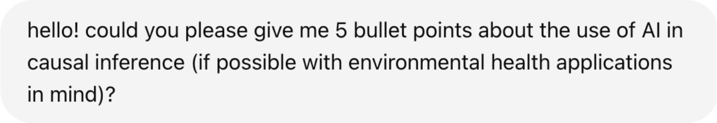
- Reproducible AI
    - Results can be sensitive to code details, random seeds, and hyperparameter settings

Note: Tools like SHAP can help make AI causal models more interpretable, but do not improve causal accuracy

AI & Causal Inference in Env Health Studies

AI & Causal Inference in Env Health Studies

- This would not be an AI talk without ChatGPT
- So I asked ChatGPT



hello! could you please give me 5 bullet points about the use of AI in causal inference (if possible with environmental health applications in mind)?

- And after some refinement of the question
  - To e.g., focus more on deep learning vs. more traditional ML methods like LASSO and BART

AI & Causal Inference in Env Health (cont'd)

- **Causal Representation Learning**:
  AI methods like deep representation learning aim to uncover latent structures that help adjust for unmeasured confounding—especially useful when linking complex, high-dimensional environmental exposures (e.g., satellite data) to health outcomes.
- **Bias Correction via Generative Models**:
  Generative AI models (e.g., variational autoencoders or GANs) can be used to impute missing environmental exposure data or simulate counterfactual exposure surfaces, improving robustness of causal effect estimates.
- **Emulation of Hypothetical Interventions**:
  Reinforcement learning and neural network-based policy learning can emulate target trials and optimize exposure reduction strategies—e.g., identifying optimal air pollution control policies to reduce adverse birth outcomes.
- **Quantifying and Propagating Exposure Uncertainty**:
  Bayesian deep learning models can estimate uncertainty in predicted exposures (e.g., PM2.5 from remote sensing) and propagate that uncertainty into health models, leading to more credible causal inference.
- **Causal Discovery in Complex Systems**:
  AI-based causal discovery algorithms (e.g., NOTEARS or DAG-GNN) can help uncover causal structures among multiple pollutants, meteorological factors, and socioeconomic variables, guiding further causal modeling in environmental health

AI & Causal Inference in Env Health (cont'd)

- There is a lot of promise on AI use in causal inference for environmental health models
- But we're definitely not there yet
- So far, more traditional ML approaches are being increasingly used
  - Both in terms of causal inference
  - And e.g., in analyzing exposure to high-dimensional mixtures in health studies
- Better adoption so far in the exposure assessment side
  - More aligned tasks to prediction
- Critical considerations that may delay further adoption

AI & Causal Inference in Env Health (cont'd)

- There is a lot of promise on AI use in causal inference for environmental health models
- But we're definitely not there yet
- So far, more traditional ML approaches are being increasingly used
  - Both in terms of causal inference
  - And e.g., in analyzing exposure to high-dimensional mixtures in health studies
- Better adoption so far in the exposure assessment side
  - More aligned tasks to prediction
- Critical considerations that may delay further adoption

In the meantime, how can we use AI
in air pollution health studies?

Current AI Use in Health Studies

1. Editing text for manuscripts/proposals
2. Bounce ideas and check concepts, e.g.,
    - Can I use AI for causal inference in air pollution health studies?
    - What are PAHs?
3. Generate graphs for presentations
4. Help with code
5. Help with literature reviews

→ Generative AI, LLMs

Help with Code

- Before gen AI, we had to spend hours googling how to code certain tasks
- Gen AI tools (e.g., chatGPT) can provide quick coding advice and examples
- Saves time and can encourage good coding practices

Help with Code

- Before gen AI, we had to spend hours googling how to code certain tasks
- Gen AI tools (e.g., chatGPT) can provide quick coding advice and examples
- Saves time and can encourage good coding practices

- However, when it comes to implementing advanced statistical methods. . .
- Gen AI can produce beautiful, albeit wrong, code very confidently

Help with Literature Reviews

- Gen AI can identify relevant papers very quickly
- And summarize findings efficiently
- Thus, greatly helping with lit reviews and saving time

Help with Literature Reviews

- Gen AI can identify relevant papers very quickly
- And summarize findings efficiently
- Thus, greatly helping with lit reviews and saving time

- Sometimes makes up papers (hallucination)
- Sometimes lacks access to full texts (e.g., paywalled journals)
- Cannot critically assess study quality or causal claims
- May produce biased or overly confident summaries

In Summary

- AI is a promising tool for causal inference in environmental health studies
    - Especially for complex, high-dimensional data
- Still in early stages, with important limitations
    - Fair, explainable, interpretable, and reproducible AI
- Causal assumptions essential
- Generalizability & transportability also concerns
- Even as currently used is very helpful (with caveats)

In Summary

- AI is a promising tool for causal inference in environmental health studies
  - Especially for complex, high-dimensional data
- Still in early stages, with important limitations
  - Fair, explainable, interpretable, and reproducible AI
- Causal assumptions essential
- Generalizability & transportability also concerns
- Even as currently used is very helpful (with caveats)

Ultimately, AI can greatly complement but not replace robust epi & stats methods training

Thank you!

Questions?

`mk3961@cumc.columbia.edu`