



Integrating Artificial Intelligence into Causal Research in Epidemiology

Ellicott C. Matthay¹ · Daniel B. Neill^{2,3,4} · Andrea R. Titus¹ · Sunita Desai⁵ · Andrea B. Troxel⁶ · Magdalena Cerdá¹ · Iván Díaz⁶ · Michele Santacatterina⁶ · Lorna E. Thorpe¹

Accepted: 4 March 2025
© The Author(s) 2025

Abstract

Purpose of Review Recent advances in Artificial Intelligence (AI) present new and not widely recognized opportunities to advance the rigor, scope, efficiency, and impact of epidemiologic research aiming to make causal inferences or causal decisions. We describe recent developments, challenges, and examples for integrating varied AI tools into the steps of Petersen and van der Laan’s causal inference roadmap and causal decision-making tasks.

Recent Findings AI tools relevant to causal research in epidemiology include predictive models, unsupervised learning, causal structure learning, causal estimation, and generative models. Opportunities exist to integrate AI at each stage of the causal roadmap. This includes the use of generative models to synthesize scientific literature and identify knowledge gaps; causal structure learning to discover or hypothesize causal structures from data; unsupervised learning from unstructured text to generate quantitative variables for analysis; predictive models to drive clinical or policy interventions; generative or causal models to assess or establish identifiability; causal models for estimating statistical parameters; and generative models to create text, tables, and figures to interpret and disseminate findings. Researchers must be mindful of potential pitfalls of AI tools such as insufficient training data, poor accuracy, biases, and ethical and legal concerns.

Summary Diverse AI tools are available to support causal research in epidemiology. Steps of the causal inference roadmap cannot yet be fully automated, but thoughtful “collaboration” between investigators and AI tools may accelerate or deepen the research at each step.

Keywords Causal inference · Artificial intelligence · Machine learning · Large language models

Introduction

Artificial Intelligence (AI) tools are rapidly being integrated into everyday technologies, including those used for epidemiologic research. AI is a sub-discipline of computer science focused on creating systems to complete tasks that usually require human intelligence, such as analyzing or generating text, images, or quantitative data. AI applications are often powered or implemented through Machine Learning (ML), processes by which computers use algorithms to analyze and learn from data and then complete specific tasks such as generating insights, predictions, or decisions. AI tools present numerous opportunities for advancing the rigor, scope, and consequences of epidemiologic research, especially research aiming to make causal inferences or causal decisions. For example, AI tools can be used to automate data collection at scale by converting unstructured text into quantitative variables for analysis [1]. In part for this reason, the number of epidemiology journal articles addressing or incorporating

✉ Ellicott C. Matthay
Ellicott.Matthay@nyulangone.org

¹ Division of Epidemiology, Department of Population Health, New York University Grossman School of Medicine, New York, NY, USA

² Courant Institute, Department of Computer Science, New York University, New York, NY, USA

³ Robert F. Wagner Graduate School of Public Service, New York University, New York, NY, USA

⁴ Center for Urban Science and Progress, Tandon School of Engineering, New York University, New York, NY, USA

⁵ Division of Healthcare Delivery Science, Department of Population Health, New York University Grossman School of Medicine, New York, NY, USA

⁶ Division of Biostatistics, Department of Population Health, New York University Grossman School of Medicine, New York, NY, USA

AI has increased markedly in recent years (Figure 1). Applications of AI relevant to the COVID-19 pandemic further fueled these advances [2].

Much has been written in the epidemiologic literature about the uses, strengths, and limitations of ML in epidemiologic research [3–14]. Researchers have highlighted the use of ML to conduct statistical modeling on large datasets with many predictor variables and complex, non-linear relationships among variables. They have also cautioned that the computer alone is not enough: such algorithms cannot be applied successfully without understanding the algorithm's assumptions, critically checking whether they are met, incorporating substantive knowledge on the variables of interest, and taking explicit steps to ensure that models do not propagate biases or discrimination.

However, AI-related discussion in the epidemiologic literature has focused almost exclusively on ML for predictive modeling in the context of estimating causal effects or heterogeneity in causal effects [3–14]. These methods constitute a narrow subset of the AI tools now widely in use. Recent advances in Generative AI, including Large Language Models (LLMs) such as ChatGPT, are widely publicized but rarely discussed in the epidemiologic literature. These broader classes of AI tools could markedly alter the conduct of epidemiologic research, for example by providing critical synthesis of the scientific literature; revealing the causal connections among variables; accelerating the pace and scale of data collection; automating components

of data processing or analysis; identifying alternative ways to achieve causal identification or conditional exchangeability; assisting with producing text, tables, or figures for scientific manuscripts; or determining the optimal combination of interventions needed to achieve a specific goal (for example, identifying which intervention(s) should be used, to whom the intervention(s) should be targeted, and how the intervention(s) should be tailored to the target population to optimally advance health equity).

The existing focus in the epidemiologic literature on ML for predictive modeling suggests that some epidemiologists may be unaware of the broader suite of AI tools now available to support epidemiologic research. To take full advantage of these advances, epidemiologists need guidance on tools available for integration into causal research, focusing both on opportunities to improve efficiency, quality, and impact and on cautions and risks. We aim to fill this gap.

This paper provides a structured overview of AI tools relevant to epidemiologic research. The term AI is loosely and ambiguously defined, and thus there exists debate on which approaches do or do not constitute AI; we do not attempt to weigh in on this debate but rather consider a wide scope of methods (ranging from statistical machine learning to deep learning with neural networks to large language models) that add value at different points in the causal inference and causal decision-making processes. As a guiding framework, we apply Petersen and van der Laan's roadmap for causal inference, [15] which traces the arc of a typical study

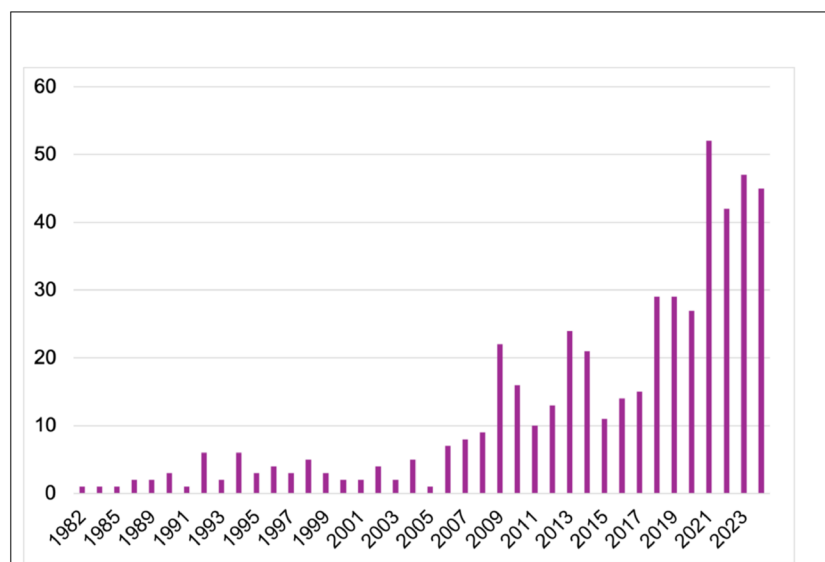


Fig. 1 Number of publications with “artificial intelligence” or related terms in the title or abstract of selected epidemiology journals, 1982–2024. Data generated from PubMed search “Results by Year” feature using the search term: (((“artificial intelligence”[Title/Abstract]) OR (“AI”[Title/Abstract]) OR (“Large language models”[Title/Abstract]) OR (“natural language processing”[Title/Abstract]) OR (“NLP”[Title/Abstract]) OR (“machine learning”[Title/Abstract]) OR (“ML”[Title/

Abstract])) AND (“American journal of epidemiology”[Journal]) OR (“International journal of epidemiology”[Journal]) OR (“Epidemiology”[Journal]) OR (“Journal of epidemiology and community health”[Journal]) OR (“Annals of epidemiology”[Journal]) OR (“European journal of epidemiology”[Journal]) OR (“Journal of clinical epidemiology”[Journal]))

aiming to estimate causal effects from refining the target causal quantity and available data through estimation and interpretation. This paper builds on work by Petersen and colleagues piloting the use of LLMs as co-pilots throughout the causal roadmap [16, 17]. Our target audience is researchers with graduate-level training in epidemiologic methods and introductory causal inference. Using applied examples throughout, we underscore the contributions AI might make, potential challenges and limitations of relying on AI, and how epidemiologists can work with AI to do better causal research. We also discuss the advancements in theory, expertise, data, and ethics that are needed for AI tools to reach their full potential for causal research in epidemiology.

A Simplified Typology of AI Tools Relevant to Causal Research in Epidemiology

No existing typology of AI tools comprehensively captures or distinguishes among tasks relevant to epidemiologists. Table 1 provides a loose classification by adapting and combining several typologies [4, 18–20]. The categories are not mutually exclusive or collectively exhaustive, but offer some intuition about key distinctions among approaches. We categorize AI tools relevant to epidemiologists into: (1) predictive models, (2) unsupervised learning, (3) causal structure learning, (4) causal estimation, and (5) generative models. For more precise definitions, see the original typologies [4, 18–20]. Also note that many methods (e.g., random forests, support vector machines, artificial neural networks) can be used for tasks in multiple categories. Among these, deep learning with artificial neural networks has gained prominence and fueled many of the most novel recent developments [21].

Integrating AI Into the Causal Roadmap

This section maps the distinct roles that AI tools can play in epidemiologic research aiming to estimate causal effects. A common framework for causal effect estimation is Petersen and van der Laan's 2014 "roadmap" [15]. Their 7-step approach—a widely used heuristic for teaching and applying causal inference concepts in epidemiology and biostatistics—traces the arc of asking and answering questions about causal effects. This includes specifying the information that is already known (e.g., a directed acyclic graph [DAG]), detailing the available data and desired causal effect (e.g., the average treatment effect of an exposure in a designated population), assessing and establishing identifiability, conducting estimation, and ultimately, interpreting the resulting statistical parameter. Opportunities exist to integrate AI into each of these steps (Table 2). The subsequent section addresses opportunities to integrate AI into causal

decision making (i.e., selecting, targeting, or tailoring an intervention).

Specifying the Causal Model and Existing Knowledge

Causal inference tasks often begin by specifying a causal model or drawing a DAG representing the investigator's assumptions about the causal structure and time ordering of relevant variables. DAGs are typically based on prior research, observed data, and expert knowledge [15]. At this step, AI may serve two purposes: First, Generative AI can synthesize the existing scientific literature and identify knowledge gaps. Second, AI tools can be used to learn or hypothesize causal structures.

Multiple LLMs, both publicly available (e.g., ChatGPT) and privately developed, are capable of synthesizing scientific literature, but to date, their reliability is poor [57–59]. LLMs learn patterns of text, speech, or language, and generate new text based on what they have learned about these patterns. However, these models do not "understand" the text's content, and they are sensitive to choice of text on which they are trained [64]. Anecdotally, public LLMs tend to produce incomplete literature reviews, provide inaccurate assessments of study quality, and fabricate both facts and references to scientific articles ("hallucinations") [57, 58]. LLMs that are trained on more complete bodies of scientific literature and that are tailored to the task of summarizing scientific literature (e.g., iris.ai [65], scite assistant [66], Stanford STORM [67]) may be more reliable, but must be validated for substantive and methodologic domains of epidemiologic research. Accuracy and completeness may also improve as this technology advances (e.g., from GPT-3 to GPT-4). If deemed adequately accurate and complete, these tools could be used to conduct targeted literature reviews to evaluate the strength of evidence for each candidate edge between two nodes in a DAG. At present, the optimal use of Generative AI may be to accelerate the work of traditional literature reviews by conducting a "first pass" that is subsequently verified by the investigator.

AI tools can also be used to learn a causal structure, such as a DAG [45]. Given a dataset and a set of assumptions (causal Markov condition, faithfulness, sufficiency, and acyclicity), the Peter and Clark (PC) algorithm [68] can be used to automate causal structure learning of Bayesian networks. PC and similar approaches identify the sets of causal structures (nodes and directed edges between them) that are consistent with the data and assumptions provided. Expert knowledge (e.g., regarding the temporal order of the variables), stronger assumptions (such as parametric model assumptions, e.g., linear relationships with non-Gaussian noise [46, 47]), or different data (e.g., on an experiment vs. observational) can be used to reduce the number of causal

Table 1 Simplified typology of AI tools in epidemiologic research

Category	Description	Examples
Predictive models	Using statistical models or algorithms to predict the value of an outcome variable based on observations of predictor variables and the outcome variable. This category includes traditional regression analysis approaches as well as more sophisticated algorithms such as artificial neural networks, support vector machines, and decision trees such as CART and random forests [22].	• Predicting future health risks for individuals or populations [23–26] • Forecasting population trends in disease and injury [27] • Tracking emerging health conditions (e.g., long COVID) when cases are under-diagnosed [28] • Estimating outcomes for small areas or populations missing data [29]
Unsupervised learning	Analyzing data that are typically large, messy, and unstructured to identify patterns. In contrast to predictive modeling, there is no outcome variable against which the models are tested or validated [30]. Unsupervised learning approaches aim to identify how the inputs (e.g., values of variables) cohere or differ across observations. These techniques include clustering (e.g., <i>k</i> -means), dimensionality reduction (e.g., latent class analysis, principal components analysis), [31, 32] density estimation, [33, 34], and anomaly and pattern detection [35–37].	• Using latent class analysis to assess how policy exposures cluster together across states and time [38, 39] • Extracting meaningful information or quantitative data from unstructured text in electronic health record notes, social media posts, or legal text [40–42]
Causal structure learning	Applying causal learning or modeling algorithms to determine causal structures among variables [19, 43, 44]	• Drawing the directed acyclic graph (DAG) or the set of possible DAGs consistent with a given dataset [45–47] • Discovering valid instrumental variables or regression discontinuity designs in large datasets [48, 49]
Causal estimation	Applying causal learning or modeling algorithms to predict counterfactuals or estimate the causal effect(s) of an intervention [19, 43, 50]	• Fitting statistical models to estimate the causal effects of public health interventions [51, 52] • Estimating heterogeneous treatment effects (i.e., variation in the causal effect of a treatment or exposure across subgroups) [53, 54] • Optimizing the targeting and tailoring of interventions [51, 52]
Generative models	Modeling the joint probability of all the observed variables and simulating new data by sampling from that distribution. “Generative” models create new data. In contrast, most predictive modeling, unsupervised learning, and causal estimation approaches are “discriminative”, because they categorize existing data or model the relationship between an outcome variable and predictor variables. The most advanced generative models use deep learning to build highly accurate generative models of complex data (e.g., natural language), trained on large amounts of data, enabling both high performance on more traditional ML tasks (e.g., text classification) and generation of novel content [55, 56].	• Synthesizing scientific literature [57–59] • Outlining or drafting text, tables, or figures for reporting scientific studies [60–62] • Applying foundation models to geospatial data to forecast and interpolate measures of socioeconomic, environmental, or health factors [63]

Table 2 Opportunities to integrate AI tools into the causal inference roadmap

Step (directly quoted from [15])	Potential roles for AI	Key challenges and limitations	Potential roles for investigator
1. Specify the knowledge about the system to be studied using a causal model. Represent background knowledge about the system to be studied. A causal model describes the set of possible data-generating processes for this system.	• Synthesize the scientific literature and identify knowledge gaps • Learn causal structures or generate candidate DAGs consistent with a given dataset and investigator inputs	• Limited accuracy (at present) • Need for human verification • Results sensitive to choice of training data • Accuracy and fairness depend on quality, biases, and accuracy of input data and input assumptions	• Define the research question • Specify basic assumptions of the setting • Review, validate, and refine scientific literature review and candidate DAGs produced by AI
2. Specify the observed data and their link to the causal model. Specify what variables have been or will be measured, and how these variables are generated by the system described by the causal model.	• Generate observed data from unstructured and messy text, images, or websites • Conduct interviews or surveys • Monitor data quality • Harmonize measures from distinct sources of data • Analyze decentralized private data	• Need for validation against human-coded “gold standards” • Balancing value of increased efficiency and larger datasets with greater potential for errors or biases • Data privacy • Reproducibility	• Determine what data are legal, ethical, and feasible to collect • Define data collection prompts for AI tools • Validate AI-generated data against human-collected reference • Assess reproducibility
3. Specify a target causal quantity. Translate the scientific question into a formal causal quantity (defined as some parameter of the distribution of counterfactual random variables). (a) On which variables will you intervene? (b) How will you set the value of these intervention variables? (c) What summary of counterfactual outcome distributions is of interest? (d) What population is of interest?	• Refine the causal question and corresponding target parameter • Apply AI-based interventions (e.g., clinical decision support tools based on individualized risk prediction algorithms)	• Protection of human subjects • Fairness	• Design AI tool to conduct the intervention (e.g., recommending specific treatments for specific patients) • Determine how to efficiently, ethically, and fairly integrate the given tool into existing health care, public health, or public policy operations
4. Assess identifiability. Assess whether it is possible to represent the target causal quantity as a parameter of the observed data distribution (estimand) and, if not, what further assumptions would allow one to do so.	• Assess whether required conditions or assumptions are met for a given causal model • Suggest alternative sets of assumptions to achieve identification • Identify valid instrumental variables in a dataset • Construct control groups for which required (but untestable) conditions or assumptions are more likely to be met	• Limited uptake and validation of some tools (e.g., automated discovery of instrumental variables from data or text; use of LLMs to suggest alternative identification approaches) in applied epidemiologic research	• Select the preferred identification strategy and corresponding set of conditions or assumptions • Assess plausibility of conditions or assumptions
5. State the statistical estimation problem. Specify the estimand and statistical model. If knowledge is sufficient to identify the causal effect of interest, commit to the corresponding estimand. If not, but one still wishes to proceed, choose an estimand that under minimal additional assumptions would equal or approximate the causal effect of interest.	• Conduct simulations to inform the selection of alternative target statistical parameters or estimation approaches • Stimulate ideas for sensitivity analyses or falsification tests	• Limited accuracy • Need for human verification • Results sensitive to choice of training data • Accuracy and fairness depend on quality, biases, and accuracy of input data and input assumptions	• Select the candidate control variables, estimators, and modeling approaches • Select appropriate sensitivity analyses

Table 2 (continued)

Step (directly quoted from [15])	Potential roles for AI	Key challenges and limitations	Potential roles for investigator
6. Estimate. Estimate the target parameter of the observed data distribution, respecting the statistical model.	• Incorporate flexible data-adaptive statistical modeling into estimators of the target causal quantity or quantities	• Obtaining valid statistical inferences (standard errors, confidence intervals) • Limited uptake • Does not solve threats to validity arising from unmeasured confounding, positivity violations, or measurement error	• Ensure statistical modeling assumptions are met
7. Interpret. Select among a hierarchy of interpretations, ranging from purely statistical to approximating a hypothetical randomized trial.	• Interpret results based on the chosen causal model, observed data, target causal parameter, and identification strategy • Assist in developing text, tables, and figures to report findings	• Limited accuracy • Need for human verification • May violate journal or funder policies	• Discuss the likelihood that identifiability conditions are met • Use knowledge of substantive area and political, organizational, and scientific context to review, validate, refine, and add nuance to reports of findings produced by AI

structures in the consistent set. Alternatively, approaches such as Fast Causal Inference [69]—an extension of the PC algorithm—can be used to find causal structures under weaker assumptions (e.g., relaxing the assumption of “no unmeasured confounders”), usually at the cost of increased runtime and a larger consistent set.

In practice, AI tools for learning causal structures have been used to evaluate the effects of gene expression on disease; [70] delineate how firearm laws, firearm ownership, and firearm mortality mutually affect one another over time; [71] and determine the temporal ordering of causal effects between depression and sleep problems [72]. Structure learning is often challenging in epidemiologic settings because of spatially and temporally correlated data. Newer approaches such as Gaussian process modeling [73] or transfer entropy in temporal data [74] may help disentangle spatiotemporal correlations from causal relationships between variables. Additionally, because any artefacts of bias or discrimination (e.g., based on race) present in the input data or input assumptions are likely to be propagated through AI algorithms, critical evaluation of the equity-related value judgements built in to the output causal structure(s) is essential [75–77].

Specifying the Observed Data and their Link to the Causal Model

The next step of the roadmap is to specify the variables that have been or will be measured and what units or participants will be observed or sampled. At this step, AI tools can support and enhance data collection and refinement of variables and measures.

Unsupervised learning tools such as natural language processing (NLP) and LLMs can generate the observed data, for example by extracting and converting unstructured text or images into quantitative variables for analysis [1]. Epidemiologic investigators are applying these methods to generate policy exposure variables from legal text; [42, 78–80] measure neighborhood environments from archived Google Street View imagery; [81] quantify neighborhood cannabis retail environments; [82, 83] measure social norms, processes, or sentiments (e.g., racism) from social media posts or mass media coverage; [41, 84, 85] identify and classify food advertisements targeting children; [86] determine social, behavioral, or clinical factors from clinical notes in Electronic Health Records (EHR); [40, 87, 88] characterize circumstances of suicides from narrative reports of medical examiners or law enforcement; [89] estimate current and future levels of air pollution exposures or disease outbreaks from historical datasets; [90–92] and track emerging disease outbreaks when cases are under diagnosed [28].

AI tools may change the scale, pace, and nature of data collection in other ways. LLMs can code themes from

interview transcripts, making it possible to complete in-depth qualitative research at quantitative scale. Recent tests of these techniques show moderate to high consistency against human-coded interviews [93–95]. Computer programs can be trained as interviewers for survey research, a practice that may reduce the risk of certain biases (e.g., social desirability) while increasing the risk of other issues (e.g., nonsensical responses) [96, 97]. LLMs may be trained to enhance questionnaire design or impute missing public opinion data, although these remain possibilities rather than established techniques [97]. Data quality monitoring must now incorporate checks to ensure that surveys intended for humans are not fraudulently completed by Generative AI programs [97]. Automation of data collection tasks also presents opportunities to iteratively refine the definitions of measures or add sensitivity analyses, since changing data collection prompts amounts to changing lines of code rather than restarting extraction from scratch.

Advancements in AI-based data harmonization and data sharing technologies present notable opportunities to create and analyze large datasets derived from separate sources. Although early in development, LLMs have been leveraged to define Common Data Elements and accelerate the process of data harmonization in biomedical research [98]. Paired with federated learning, an AI technique that permits separate groups to contribute private decentralized data to train a single centralized model, these tools can facilitate the compilation and analysis of diverse measures across distinct health systems or cohorts while protecting private sensitive information [99]. These advancements show promise, for example, in assessing nationwide clinical outcomes among organ transplant recipients [99]. AI tools for harmonizing data across separate sources may also enhance internal validity by facilitating adjustment for confounders only available across distinct datasets, and external validity by incorporating separate cohorts to increase population representativeness. More recently, researchers have also applied foundation models—another type of generative AI—to large streams of geo-indexed data to achieve state-of-the-art performance for forecasting and interpolating county-level health, environmental, and socioeconomic indicators.

AI-driven paradigm shifts in epidemiologic data collection pose new ethical questions and challenges [97]. Propagation of biases or discrimination in training data are a threat to fairness and validity [97]. Predicting individuals' opinions may raise new questions about participant consent. AI-based data collection powered by proprietary algorithms may hinder reproducibility. Interview transcripts and patient notes in electronic health records contain identifiers and private information protected by the Health Insurance Portability and Accountability Act (HIPAA) and therefore cannot legally be input into public LLMs. Investigators must

therefore proceed using private, HIPAA-compliant LLMs, or human- or AI-driven de-identification [100].

Specifying the Target Causal Quantity

The third step is to define the research question as a formal quantity or parameter corresponding to the causal effect of a specific intervention or exposure on an outcome variable in a defined target population. It may be possible to train LLMs to select or define causal parameters of interest. For instance, given a causal model and a research question, a generative AI tool could instruct the investigator on which target causal quantities are identifiable, and which of those best reflects the original research question.

Another salient use of AI at this stage is as the intervention itself. As an intervention, an AI tool might determine what exposure variable to intervene on and how to modify the chosen exposure. The research then aims to infer its causal effect of this system on the outcome in the target population. For example, clinical researchers have applied predictive models to EHR data to stratify patients based on their risk of cancer recurrence, sepsis, post-surgery complications, or high utilization of healthcare resources, and used these risk predictions to provide tailored support for clinician decision-making in caring for each patient (i.e., applying dynamic treatment rules) [101–107]. The target causal quantity could then be the average level of the outcome had all eligible patients been exposed to the decision support tool compared with the average outcome had all eligible patients *not* been exposed to the tool (i.e., an average treatment effect), but other summaries of the counterfactual outcome distributions based on subgroups or effect modifiers may also be of interest.

Beyond healthcare, public policy interventions involving AI may also be exposures of interest to epidemiologists. For example, the US Department of Justice invested in research to evaluate the use of ML to predict the future recidivism risk among individuals released from prison to parole and to tailor programming accordingly [108]. As with data collection, AI-based interventions present opportunities to reduce certain biases, for example by reducing interpersonal racial discrimination in sentencing, but may increase the risk of other concerns, including ethical or safety risks to patients or parolees if decision support tools fail and risks of propagating biases and discrimination [109].

Assessing and Establishing Identifiability

Assessing identifiability means determining, for a given causal model (e.g., DAG) and target causal quantity, whether the measured variables and observations are sufficient to meet the required conditions [15, 110]. This typically means ensuring that all confounders have been correctly

identified, measured, and controlled, or that there exists a valid instrumental variable that can be leveraged to make causal inferences [111]. AI tools are available to determine whether these criteria are met, and to select or generate control groups or datasets that are more likely to meet these (untestable) criteria.

For a given causal model with variables designated as measured or unmeasured, simple automated software such as DAGitty can readily identify sets of variables sufficient to control confounding [112]. ML has also been used to build control groups or counterfactuals. For example, synthetic control methods create artificial control groups by taking weighted averages of the outcomes in untreated units, with the weights selected algorithmically by minimizing differences in confounder values between the treated and synthetic control units [113, 114]. These methods have been used in epidemiologic research to estimate the causal effects of a variety of public policies. ML algorithms for automated discovery of valid instruments in large datasets have been developed to identify local average treatment effects, [48, 49] although these methods have not yet been applied in epidemiologic research.

Theoretically, LLMs could be trained to analyze news media, proposed bills, legislation, regulations, or legal documents to identify new opportunities for quasi-experiments, for example if a new public policy were rolled out via lottery. We are not aware of any existing applications of this approach. Further, in situations where there is uncertainty about whether the assumptions required for identification are met, LLMs could be used to simulate datasets or causal models under alternative scenarios to determine how identification could be achieved under each scenario. As with literature reviews, human verification of the accuracy of LLM output is essential.

Stating the Statistical Estimation Problem

At the fifth step, researchers must specify the statistical model to be used to estimate the target causal quantity and determine whether the observed data are adequate to estimate the target. If so, the study can proceed with estimation. If not, the target must be altered or the set of assumptions expanded (i.e., return to the previous step). AI tools to support these decisions are similar to those described for assessing and establishing identifiability. More broadly, at this stage, LLMs can support analytic decision-making by synthesizing recommended approaches in the scientific literature or supporting simulations to guide the choice of analytic modifications. For example, Generative AI could be used to simulate complex, realistic datasets with known parameters for the researcher to use to select among alternative estimation approaches [115]. The identified optimal approach can then be applied to the real data. Similarly, interactions with

Generative AI can stimulate ideas for sensitivity or falsification analyses. As with step 1, Generative AI should only be used to synthesize information or make recommendations if the accuracy and quality of the output can be verified.

Estimating the Target Causal Quantity

The most common application of AI in epidemiologic research is the use of semi-parametric modeling techniques when estimating causal effects [4, 15, 116]. For example, random forests, artificial neural networks, support vector machines, or a combination of these might be used to model the outcome variable as a function of the exposure and confounders, replacing traditional parametric regressions. These approaches are advantageous because they allow for data-driven model selection, flexible shapes of the relationships between variables, many predictor variables, and complex interactions among predictors [117–119]. Recent applications of deep learning may further optimize the task of causal estimation by automating the selection of estimators across a vast array of possible data structures and statistical procedures [120] or automating the derivation of formulas to compute standard errors [121]. All of these tools are relevant to analyses estimating average treatment effects as well as causal mediation and transportability analyses [122–124].

Substantial progress and attention have also been directed to ML tools for estimating heterogeneous treatment effects (HTEs). Here, multiple distinct tasks are relevant, including data-driven identification of subgroups that respond differently to [53, 54] or benefit most from [125] an intervention, or testing for heterogeneity across all covariate subgroups [126]. For example, researchers used causal forest modeling to identify subgroups of randomized trial participants who benefitted most from an intensive weight loss intervention, according to their HbA_{1c} and self-reported general health at baseline [127].

Obtaining valid statistical inferences from statistical models that incorporate ML can be a challenge, because there is limited statistical theory on which to base the estimation of standard errors or confidence intervals [128]. Targeted Maximum Likelihood Estimation (TMLE), debiased machine learning, and balancing estimators are exceptions [129–133]. Among these, TMLE has gained distinction in epidemiologic research [134].

Interpreting and Reporting Results

Once the target causal parameter has been estimated, the results must be interpreted and reported appropriately. At this stage, generative AI may assist in selecting among the possible levels of interpretation, ranging from a statistical parameter of the observed data to an effect that approximates that from a randomized trial [15]. As the strength of the

interpretation depends on the likelihood that the identifiability conditions are met, ML algorithms for learning causal structures and establishing identification may also come into play here (see step 4 above).

Generative AI may also support investigators in creating the text, tables, and figures for scientific manuscripts and presentations reporting study findings [60, 135]. For these tasks, AI tools may be best thought of as a collaborator in the writing process, rather than a stand-alone production tool. For example, LLMs can support the efficiency or quality of writing by developing outlines or first drafts or providing feedback on grammar or logical arguments. Studies testing AI for scientific and medical writing have found that ChatGPT can improve readability of abstracts and introduction sections compared with human-generated text, but in some cases the content quality was inferior, [61, 62] emphasizing the need for adequate investigator oversight to prevent biases and inaccuracies [60, 136]. One large evaluation of LLM-generated feedback on research papers found substantial overlap between LLM and human feedback, demonstrating the potential utility of LLMs as a complement to expert feedback [137]. Incorporating LLMs into the writing process may also have beneficial effects on equity in scientific fields by reducing barriers experienced by non-native English speakers. Other opportunities for enhancing dissemination of research findings, for example creating eighth-grade reading level summaries, [138] continue to be explored.

Most funders, journals, and publishers now have policies regarding the use of AI in scientific writing [139]. For example, the Journal of the American Medical Association (JAMA) discourages but does not ban the use of AI-generated content, and requires that authors disclose how AI was used in the study's conduct and reporting [140]. Generative AI models cannot generally be considered authors because they cannot be held responsible for a manuscript's contents [140]. The National Institutes of Health permits the use of Generative AI to assist in grant writing, but bans its use in peer review [141]. There are currently few practical ways to enforce these bans but, as with the technology itself, this could change rapidly [142]. Because LLMs can memorize and regurgitate their training data, the risk of plagiarism may be substantial [143]. It is therefore wise to run all drafts through plagiarism detection software.

Integrating AI into Causal Decision-making

Beyond causal effect estimation, epidemiologists also aim to *make causal decisions*—for example, determining which intervention(s) should be used, to whom the intervention(s) should be targeted, or how the intervention(s) should be tailored to the target population [144, 145]. This task is distinct from causal effect estimation because the quantity of interest is the intervention

assignment itself, not the effect of the intervention on the outcome [144]. However, AI tools similarly present multiple opportunities to enhance causal decision-making tasks.

Given a set of candidate interventions, estimates of their causal effects on a health outcome of interest for relevant populations, and chosen constraints (e.g., budget, fairness), predictive modeling and optimization approaches can be used to determine which intervention(s) will achieve a specific goal, for example maximally reducing the given outcome in the overall population [146–149]. For example, one study applying this approach concluded that designated targets for reducing overdose deaths in the US are only possible if broader availability of medication treatment for opioid use disorder is paired with increased distribution of the overdose reversal agent naloxone, but not if either policy is enacted alone [51]. The optimization may also include constraints or penalties designed to improve fairness or reduce disparities between groups, to ensure that the benefits of an intervention are more equitably distributed across the population.

AI tools for causal decision-making can also support geographic targeting of an intervention to areas where it is most likely to be most effective. For example, researchers have applied predictive modeling to anticipate where burden will be highest and to dynamically adapt where resources are targeted in response. The PROVIDENT trial is testing this approach to anticipate and prevent local surges in drug overdoses [52, 150, 151]. Similarly, predictive models that identify subgroups who benefit most from an intervention can be used to determine to whom an intervention should be targeted [152]. Statistical methods for transporting or generalizing causal effect estimates can also incorporate predictive modeling and inform targeting efforts by estimating the potential impact of an intervention in a novel target population that differs in composition from the original study population [153, 154].

AI tools may also be leveraged to inform the tailoring of intervention(s) to each individual according to their baseline characteristics or responses to the intervention(s) over the course of the study, as in the case of estimating optimal dynamic treatment rules [155]. For example, investigators have applied AI algorithms to identify which justice-involved adults would most-benefit from cognitive behavioral therapy to reduce criminal-reoffending [156]. These tasks are similar in nature to those used to estimate HTEs.

Importantly, applications of AI—particularly those involving risk-based targeting or tailoring of interventions—can perpetuate harmful stereotypes and discrimination based on race, ethnicity, gender, ability, and other social statuses [157, 158]. For example, Obermeyer found evidence of racial bias in one AI algorithm widely used in US health care, such that the algorithm assigned the same level of risk to sicker Black patients as to healthier white patients [159].

Because of the potential harms arising from discriminatory algorithms, transparent and structured evaluations of fairness—ideally led by members of minoritized groups and individuals with lived experience of marginalization—must be incorporated into the design and application of causal decision-making algorithms. For example, AI analyses grounded in the epidemiologic concept of “allowable” covariates and the inherent value judgements in covariate selection may be better positioned to prevent or mitigate AI bias [76]. Additionally, investigations supported by sociological theory or frameworks underlying the relationship between the intervention and health inequities may be better positioned to identify and disrupt rather than reinforce inequities [160]. In response to concerns about AI bias, the National Institute on Minority Health and Health Disparities developed the Science Collaborative for Health disparities and Artificial intelligence bias Reduction (ScHARE) platform [161]. ScHARE provides a low-cost collaborative cloud computing platform and access to big datasets on social determinants and health care outcomes with the goals of increasing participation of underrepresented groups in AI science and mitigating AI bias in health research. Overall, fairness in AI is a rapidly evolving area of research in computer science, bioethics, and related fields, and advancements in this area will likely have important implications for epidemiologic research involving causal effect estimation and causal decision-making.

Discussion

We provide a structured summary of opportunities to integrate recent advances in AI into causal inference and causal decision-making in epidemiology. Along the arc of a causal epidemiologic research project, AI tools for prediction, unsupervised learning, causal structure learning, causal estimation, and content generation may enhance the scale, complexity, efficiency, or quality of the research. Yet substantial limitations in accuracy, fairness, ethics, and safety remain. AI cannot yet be used to automate the scientific process; human experts remain the foundation of sound epidemiologic research. However, when viewed as an assistant in the process of conducting causal research, AI presents opportunities for thoughtful “collaboration”. AI brings new tools, but the major goals, processes, and requirements of causal research in epidemiology remain unchanged [162].

To leverage the full potential of AI in research, epidemiologists must build interdisciplinary partnerships, develop tailored data and computational resources, and navigate ethical considerations. Teams aiming to rigorously incorporate AI into epidemiologic research will benefit from interdisciplinary expertise in subdisciplines of computer science including AI and data science, statistics, machine learning, bioinformatics, medical and research ethics, and relevant

clinical or substantive areas. Clear communication of the uses and outputs of a given AI tool is also essential, because perceived “black boxes” are less likely to be trusted or used by researchers, practitioners, or the public. In our experience, effective collaborations require introducing causal inference concepts to computer and data scientists, introducing AI concepts to epidemiologists, and developing tools tailored to epidemiologic research.

Data present concerns in at least three regards. First, the types of epidemiologic data best suited to AI applications are not yet established. Second, AI algorithms depend on the quality and completeness of the data used to train them [2]. Gaps in the data necessary for appropriate training of AI models—for example because of publication bias—will limit the utility of the resulting tools. Third, the evolution of many AI technologies has gone hand-in-hand with the increasing availability of large, high-dimensional datasets. Many of these big datasets come with their own issues and biases. Because the use of AI in epidemiology intersects with big data, there are overlaid technical and analytical challenges to causal inference in the combined context of big data and AI that must be addressed in concert. For example, AI models run on big data can be extremely computationally intensive. Taking full advantage of AI therefore often requires proficiency and resources in cloud computing and super computers. Changes in the distribution of big data over time, or differences among data collected in different jurisdictions, may limit the generalizability of causal inferences and the quality of estimation.

Generative AI should only be used to synthesize information or make recommendations if the accuracy and quality of the output can be verified. This raises questions about *how* researchers should evaluate performance or accuracy. One large systematic review of this topic found that current practices for evaluation are varied, limited, and unstandardized, and consequently proposed a framework for standardizing human evaluation of LLMs in healthcare [163]. Similar investigations and standardization of practices for incorporating AI into epidemiologic research may help promote more rigorous use of these tools.

Much has been written about ethical and safety concerns with AI in medical and public health research [96, 157, 158, 164–168]. Concerns about health information privacy, data management, and data sharing are prevalent, as public LLMs such as ChatGPT cannot be used with information protected by the Health Insurance Protection and Accountability Act (HIPAA). Our institution, NYU Langone Health, is one of few academic medical centers with a private, internal, HIPAA-compliant instance of GPT-4 for use in scientific research and clinical care. The prior section discussed issues of bias and discrimination. AI tools also present novel threats to study participant safety, biosecurity, and biosafety that may require institutional review boards to learn about new technologies and adapt accordingly. Epidemiologist will have an important role to play in research

supporting the regulation of AI in health care and biomedicine, for example by conducting post-market monitoring of AI-enabled medical devices [169]. Questions about intellectual property, peer review, and replication arise when generative AI is used to create research products, making it important to monitor evolving regulations and avoid plagiarism and policy violations [139, 142, 164]. Greater consensus is needed on what responsibilities researchers have for judging the quality of AI-assisted research and managing the societal implications of using AI models that may be “overconfident”, inadvertently cause harm, or reduce public trust in science [158]. The field of epidemiology will need to wrestle with these questions, and would likely benefit from developing trainings for doctoral and post-doctoral investigators in the responsible use of AI.

We note some limitations of this review. First, AI and research on its uses in science are evolving rapidly. We aim to illustrate potential uses and considerations for AI within causal frameworks, but the applications, opportunities, and limitations described here are not exhaustive. Second, we describe opportunities to fit AI into causal frameworks, but the relation between causal inference and AI is in fact bidirectional, with opportunities to inject causal thinking into AI frameworks as well (see for example [170]). Finally, we focus on applications to causal research in epidemiology, as there is unique complementarity and rapid advancement happening at the intersection of AI and causal inference, but AI may also be useful for other undertakings fulfilled by epidemiologists including descriptive epidemiology, evidence synthesis, and implementation research.

Rigorous epidemiologic research incorporating AI can advance causal inference and causal decision-making in public health. Many AI tools are underutilized but poised to boost the innovation, efficiency, and scope of epidemiology research if applied thoughtfully and ethically with wariness of potential pitfalls. Causal research in epidemiology cannot yet be automated but anticipating which components of the research process are likely to be rigorously automated soon will facilitate long-term planning for the evolution of epidemiologic research as a field.

Key References

- Petersen M, Alaa A, Kıcıman E, Holmes C, van der Laan M. Artificial Intelligence–Based Copilots to Generate Causal Evidence. *NEJM AI*. 2024;1:AIp2400727.
 - This paper discusses the use of large language models as assistants at each step of the causal inference roadmap. The current paper builds on this foundational work.
- Gwon YN, Kim JH, Chung HS, Jung EJ, Chun J, Lee S, et al. The Use of Generative AI for Scientific Literature

Searches for Systematic Reviews: ChatGPT and Microsoft Bing AI Performance Evaluation. *JMIR Med Inform*. 2024;12:51187.

- Findings from this validation study indicate that common generative AI tools may not produce high quality scientific literature reviews.
- Chen RJ, Wang JJ, Williamson DFK, Chen TY, Lipkova J, Lu MY, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat Biomed Eng*. 2023;7:719–42.
 - This perspective paper reviews concepts of fairness in machine learning, how algorithmic biases arise in health-related research, and potential tools for mitigating these biases.
- Mathis WS, Zhao S, Pratt N, Weleff J, De Paoli S. Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: How does it compare to traditional methods? *Computer Methods and Programs in Biomedicine*. 2024;255:108356.
 - Findings from this study demonstrate proof-of-concept that large language models can be used to conduct thematic coding of qualitative interviews, enabling qualitative-depth research at quantitative scale.
- Liang W, Zhang Y, Cao H, Wang B, Ding DY, Yang X, et al. Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis. *NEJM AI*. 2024;1:AIoa2400196.
 - This validation study provides proof-of-concept that large language models can provide legitimate feedback on scientific writing. This paper is an example of the type of validation work currently being done to test the use of AI tools in scientific writing.

Author Contribution E.C.M. wrote the main manuscript text. All authors critically reviewed the manuscript and made additions.

Funding No funding was received to assist with the preparation of this manuscript.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing Interests The authors declare no competing interests.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

- VoPham T, Hart JE, Laden F, Chiang Y-Y. Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology. *Environ Health*. 2018;17:40.
- Chen J, See KC. Artificial Intelligence for COVID-19: Rapid Review. *J Med Internet Res*. 2020;22:e21476.
- Wiemken TL, Kelley RR. Machine Learning in Epidemiology and Health Outcomes Research. *Annu Rev Public Health*. 2020;41:21–36.
- Bi Q, Goodman KE, Kaminsky J, Lessler J. What is Machine Learning? A Primer for the Epidemiologist. *Am J Epidemiol*. 2019;188:2222–39.
- Bellinger C, Mohamed Jabbar MS, Zaiane O, Osornio-Vargas A. A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health*. 2017;17:907.
- Morgenstern JD, Rosella LC, Costa AP, de Souza RJ, Anderson LN. Perspective: Big Data and Machine Learning Could Help Advance Nutritional Epidemiology. *Adv Nutr*. 2021;12:621–31.
- Jorm LR. Commentary: Towards machine learning-enabled epidemiology. *Int J Epidemiol*. 2020;49:1770–3.
- Russo S, Bonassi S. Prospects and Pitfalls of Machine Learning in Nutritional Epidemiology. *Nutrients*. 2022;14:1705.
- Scheinker D, Valencia A, Rodriguez F. Identification of Factors Associated With Variation in US County-Level Obesity Prevalence Rates Using Epidemiologic vs Machine Learning Models. *JAMA Netw Open*. 2019;2:e192884.
- Kreatsoulas C, Subramanian SV. Machine learning in social epidemiology: Learning from experience. *SSM Popul Health*. 2018;4:347–9.
- Inoue K. Causal inference and machine learning in endocrine epidemiology. *Endocr J*. 2024;71(10):945–53.
- Sung J, Hopper JL. Co-evolution of epidemiology and artificial intelligence: challenges and opportunities. *Int J Epidemiol*. 2023;52:969–73.
- Broadbent A, Grote T. Can Robots Do Epidemiology? Machine Learning, Causal Inference, and Predicting the Outcomes of Public Health Interventions. *Philos Technol*. 2022;35:14.
- Hamilton AJ, Strauss AT, Martinez DA, Hinson JS, Levin S, Lin G, et al. Machine learning and artificial intelligence: applications in healthcare epidemiology. *Antimicrob Steward Healthc Epidemiol*. 2021;1:e28.
- Petersen ML, van der Laan MJ. Causal Models and Learning from Data: Integrating Causal Modeling and Statistical Estimation. *Epidemiology*. 2014;25:418.
- Alaa A, Phillips RV, Kıcıman E, Balzer LB, Laan M van der, Petersen M. Large Language Models as Co-Pilots for Causal Inference in Medical Studies [Internet]. *arXiv*; 2024 [cited 2025 Jan 8]. Available from: <http://arxiv.org/abs/2407.19118>
- Petersen M, Alaa A, Kıcıman E, Holmes C, van der Laan M. Artificial intelligence-based copilots to generate causal evidence. *NEJM AI*. 2024;1(12):AIP2400727. <https://doi.org/10.1056/AIP2400727>.
- Burkov A. The hundred-page machine learning book [Internet]. Quebec City, QC, Canada; 2019 [cited 2024 Aug 1]. Available from: <https://cir.nii.ac.jp/crid/1130282269115099008>
- Leist AK, Klee M, Kim JH, Rehkopf DH, Borda SPA, Muniz-Terrera G, et al. Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences. *Sci Adv*. 2022;8:eabk1942.
- Jebara T. Machine learning: discriminative and generative. Springer Science & Business Media; 2012. https://www.google.com/books/edition/Machine_Learning/g5rSBwAAQB-AJ?hl=en&gbpv=0.
- Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT Press; 2016.
- Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2009.
- Morgenstern JD, Buajitti E, O'Neill M, Piggott T, Goel V, Fridman D, et al. Predicting population health with machine learning: a scoping review. *BMJ Open*. 2020;10:e037860.
- Patel B, Sengupta P. Machine learning for predicting cardiac events: what does the future hold? *Exp Rev Cardiovasc Ther*. 2020;18:77–84.
- Stark GF, Hart GR, Nartowt BJ, Deng J. Predicting breast cancer risk using personal health data and machine learning models. *PLOS ONE*. 2019;14:e0226765.
- Boudreaux ED, Rundensteiner E, Liu F, Wang B, Larkin C, Agu E, et al. Applying Machine Learning Approaches to Suicide Prediction Using Healthcare Data: Overview and Future Directions. *Front Psychiatry* [Internet]. 2021 [cited 2024 Oct 28];12. Available from: <https://www.frontiersin.org/journals/psychiatry/articles/https://doi.org/10.3389/fpsy.2021.707916/full>
- Colson KE, Rudolph KE, Zimmerman SC, Goin DE, Stuart EA, van der Laan M, et al. Optimizing matching and analysis combinations for estimating causal effects. *Sci Rep*. 2016;6:23222.
- Pfaff ER, Girvin AT, Bennett TD, Bhatia A, Brooks IM, Deer RR, et al. Identifying who has long COVID in the USA: a machine learning approach using N3C data. *Lancet Digit Health*. 2022;4:e532–41.
- Viljanen M, Meijerink L, Zwakhals L, van de Kastelee J. A machine learning approach to small area estimation: predicting the health, housing and well-being of the population of Netherlands. *Int J Health Geogr*. 2022;21:4.
- Fukunaga K. Introduction to statistical pattern recognition [Internet]. Boston : Academic Press; 1990 [cited 2024 Nov 26]. Available from: <http://archive.org/details/introductiontos1990fuku>
- Rokach L, Maimon O. Clustering Methods. In: Maimon O, Rokach L, editors. *Data Mining and Knowledge Discovery Handbook* [Internet]. Boston, MA: Springer US; 2005 [cited 2024 Oct 28]. p. 321–52. Available from: https://doi.org/10.1007/0-387-25465-X_15

32. Nylund-Gibson K, Choi AY. Ten frequently asked questions about latent class analysis. *Transl Iss Psychol Sci*. 2018;4:440–61.
33. Silverman BW. Density estimation for statistics and data analysis [Internet]. Chapman & Hall/CRC; 1998 [cited 2024 Nov 26]. Available from: http://archive.org/details/densityestimation00silv_0
34. Scott DW. Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons; 2015. https://www.google.com/books/edition/Multivariate_Density_Estimation/plAZBwAAQBAJ?hl=en&gbpv=0.
35. Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. *ACM Comput Surv*. 2009;41:1–58.
36. Neill DB. Fast Subset Scan for Spatial Pattern Detection. *J R Stat Soc Ser B: Stat Methodol*. 2012;74:337–60.
37. McFowland E, Speakman S, Neill DB. Fast Generalized Subset Scan for anomalous pattern detection. *J Mach Learn Res*. 2013;14:1533–61.
38. Erickson DJ, Lenk KM, Toomey TL, Nelson TF, Jones-Webb R. The alcohol policy environment, enforcement and consumption in the United States. *Drug Alcohol Rev*. 2016;35:6–12.
39. Matthay EC, Gottlieb LM, Rehkopf D, Tan ML, Vlahov D, Glymour MM. What to Do When Everything Happens at Once: Analytic Approaches to Estimate the Health Effects of Co-Occurring Social Policies. *Epidemiol Rev*. 2022;43:33–47.
40. Patra BG, Sharma MM, Vekaria V, Adekkanattu P, Patterson OV, Glicksberg B, et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *J Am Med Inform Assoc*. 2021;28:2716–27.
41. Nguyen TT, Meng H-W, Sandeep S, McCullough M, Yu W, Lau Y, et al. Twitter-derived measures of sentiment towards minorities (2015–2016) and associations with low birth weight and preterm birth in the United States. *Comput Human Behav*. 2018;89:308–15.
42. Nay J. Natural Language Processing and Machine Learning for Law and Policy Texts. In: Katz DM, Dolin R, Bommarito M, editors. *Legal Informatics* [Internet]. Cambridge University Press; 2018 [cited 2024 Nov 26]. Available from: <https://papers.ssrn.com/abstract=3438276>
43. Cox LA. Toward practical causal epidemiology. *Global. Epidemiology*. 2021;3:100065.
44. Hünernmund P, Bareinboim E. Causal inference and data fusion in econometrics. *Econom J*. 2025;28(1):41–82. <https://doi.org/10.1093/ectj/utad008>.
45. Scutari M, Denis JB. Bayesian networks: with examples in R. 2nd ed. Chapman and Hall/CRC; 2021. <https://doi.org/10.1201/9780429347436>.
46. Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *J Mach Learn Res*. 2006;7:2003–30.
47. Shimizu S. Lingam: Non-Gaussian Methods for Estimating Causal Structures. *Behaviormetrika*. 2014;41:65–98.
48. Herlands W, McFowland III E, Wilson AG, Neill DB. Automated Local Regression Discontinuity Design Discovery. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* [Internet]. New York, NY, USA: Association for Computing Machinery; 2018 [cited 2024 Oct 9]. p. 1512–20. Available from: <https://dl.acm.org/doi/https://doi.org/10.1145/3219819.3219982>
49. Jakubowski B, Somanchi S, Iii EM, Neill DB. Exploiting Discovered Regression Discontinuities to Debias Conditioned-on-observable Estimators. *J Mach Learn Res*. 2023;24:1–57.
50. Herlands W, Neill DB, Nickisch H, Wilson AG. Change Surfaces for Expressive Multidimensional Change-points and Counterfactual Prediction. *J Mach Learn Res*. 2019;20:1–51.
51. Cerdá M, Hamilton AD, Hyder A, Rutherford C, Bobashev G, Epstein JM, et al. Simulating the simultaneous impact of medication for opioid use disorder and naloxone on opioid overdose death in eight New York counties. *Epidemiology* [Internet]. 2024 [cited 2024 Mar 12]; Available from: https://journals.lww.com/epidem/abstract/9900/simulating_the_simultaneous_impact_of_medication.224.aspx
52. Marshall BDL, Alexander-Scott N, Yedinak JL, Hallowell BD, Goedel WC, Allen B, et al. Preventing Overdose Using Information and Data from the Environment (PROVIDENT): protocol for a randomized, population-based, community intervention trial. *Addiction*. 2022;117:1152–62.
53. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci*. 2016;113:7353–60.
54. Wager S, Athey S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *J Am Stat Assoc*. 2018;113:1228–42.
55. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. *Advances in Neural Information Processing Systems* [Internet]. Curran Associates, Inc.; 2017 [cited 2024 Dec 30]. Available from: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
56. OpenAI. GPT-4 Technical Report. 2023.
57. Gwon YN, Kim JH, Chung HS, Jung EJ, Chun J, Lee S, et al. The Use of Generative AI for Scientific Literature Searches for Systematic Reviews: ChatGPT and Microsoft Bing AI Performance Evaluation. *JMIR Med Inform*. 2024;12:51187.
58. Luo X, Chen F, Zhu D, Wang L, Wang Z, Liu H, et al. Potential Roles of Large Language Models in the Production of Systematic Reviews and Meta-Analyses. *J Med Internet Res*. 2024;26:e56780.
59. Zhang G, Jin Q, Jered McInerney D, Chen Y, Wang F, Cole CL, et al. Leveraging generative AI for clinical evidence synthesis needs to ensure trustworthiness. *J Biomed Inform*. 2024;153:104640.
60. Shopovski J. Generative Artificial Intelligence, AI for Scientific Writing: A Literature Review [Internet]. Preprints; 2024 [cited 2024 Aug 14]. Available from: <https://www.preprints.org/manuscript/202406.0011/v1>
61. Sikander B, Baker JJ, Deveci CD, Lund L, Rosenberg J. ChatGPT-4 and Human Researchers Are Equal in Writing Scientific Introduction Sections: A Blinded, Randomized Non-inferiority Controlled Study. *Cureus*. 2023;15:e49019.
62. Hwang T, Aggarwal N, Khan PZ, Roberts T, Mahmood A, Griffiths MM, et al. Can ChatGPT assist authors with abstract writing in medical journals? Evaluating the quality of scientific abstracts generated by ChatGPT and original abstracts. *PLOS ONE*. 2024;19:e0297701.
63. Agarwal M, Sun M, Kamath C, Muslim A, Sarker P, Paul J, et al. General Geospatial Inference with a Population Dynamics Foundation Model [Internet]. arXiv; 2024 [cited 2025 Jan 8]. Available from: <http://arxiv.org/abs/2411.07207>
64. González J, Nori AV. Does Reasoning Emerge? Examining the Probabilities of Causation in Large Language Models [Internet]. arXiv; 2024 [cited 2024 Oct 10]. Available from: <http://arxiv.org/abs/2408.08210>
65. Dezea. Iris.ai [Internet]. Iris.ai. 2024 [cited 2024 Nov 26]. Available from: <https://iris.ai>
66. Scite. Your AI Research Assistant [Internet]. scite.ai. 2024 [cited 2024 Nov 26]. Available from: <https://scite.ai>
67. Stanford Open Virtual Assistant Lab. STORM [Internet]. 2024 [cited 2024 Nov 26]. Available from: <https://storm.genie.stanford.edu/>

68. Spirtes P, Glymour C. An Algorithm for Fast Recovery of Sparse Causal Graphs. *Soc Sci Comput Rev.* 1991;9:62–72.
69. Spirtes PL, Meek C, Richardson TS. Causal inference in the presence of latent variables and selection bias. *arXiv:1302.4983*; 2013.
70. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, GuhaThakurta D, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet.* 2005;37:710–7.
71. Barak Ventura R, Macinko J, Ruiz Marín M, Porfiri M. Association of State Firearm Laws With Firearm Ownership and Mortality. *AJPM Focus.* 2024;3:100250.
72. Rosenström T, Jokela M, Puttonen S, Hintsanen M, Pulkki-Råback L, Viikari JS, et al. Pairwise Measures of Causal Direction in the Epidemiology of Sleep Problems and Depression. *PLOS ONE.* 2012;7:e50841.
73. Flaxman SR, Neill DB, Smola AJ. Gaussian Processes for Independence Tests with Non-iid Data in Causal Inference. *ACM Trans Intell Syst Technol.* 2015;7(2):1–23.
74. Porfiri M, Ruiz Marín M. Transfer entropy on symbolic recurrences. *Chaos: Interdiscipl J Nonlinear Sci.* 2019;29:063123.
75. Chen RJ, Wang JJ, Williamson DFK, Chen TY, Lipkova J, Lu MY, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat Biomed Eng.* 2023;7:719–42.
76. Jackson JW. Meaningful Causal Decompositions in Health Equity Research: Definition, Identification, and Estimation Through a Weighting Framework. *Epidemiology.* 2021;32:282.
77. Chang T-H, Nguyen TQ, Jackson JW. The Importance of Equity Value Judgments and Estimator-Estimand Alignment in Measuring Disparity and Identifying Targets to Reduce Disparity. *American Journal of Epidemiology.* 2024;193:536–47.
78. Burris S. Building the Discipline of Policy Surveillance: Report and Next Steps from an International Convening [Internet]. Rochester, NY: Social Science Research Network; 2018 [cited 2024 Nov 27]. Available from: <https://papers.ssrn.com/abstract=3218420>
79. Korostelev M. Award # 1746192 - SBIR Phase I: Machine Assisted Comparative Policy Analysis in Public Health [Internet]. U.S. National Science Foundation Award Search. 2018 [cited 2024 Nov 27]. Available from: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1746192&HistoricalAwards=false
80. Rothenberg J, Smith L, Rencken C, Ruben E, Rowhani-Rahbar A, Smart R. Challenges and best practices in surveying gun laws over time: examining gun-free zone legislation as a case study. In: 2024 national research conference for the prevention of firearm-related harms. Seattle, WA; 2024.
81. Rundle AG, Bader MDM, Mooney SJ. Machine Learning Approaches for Measuring Neighborhood Environments in Epidemiologic Studies. *Curr Epidemiol Rep.* 2022;9:175–82.
82. Matthey EC, Gupta A, Mousli L, Schmidt LA. Using Online Crowdsourced Data to Measure the Availability of Cannabis Home Delivery: A Pilot Study. *J Stud Alcohol Drugs.* 2023;84:330–4.
83. Matthey EC, Mousli L, Ponicki WR, Glymour MM, Apollonio DE, Schmidt LA, et al. A Spatiotemporal Analysis of the Association of California City and County Cannabis Policies with Cannabis Outlet Densities. *Epidemiology.* 2022;33:715–25.
84. Relia K, Akbari M, Duncan D, Chunara R. Socio-spatial Self-organizing Maps: Using Social Media to Assess Relevant Geographies for Exposure to Social Processes. *Proc ACM Hum-Comput Interact.* 2018;2:1–23.
85. Choi S, Lee J, Kang M-G, Min H, Chang Y-S, Yoon S. Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks. *Methods.* 2017;129:50–9.
86. Palmer G, Green M, Boyland E, Vasconcelos YSR, Savani R, Singleton A. A deep learning approach to identify unhealthy advertisements in street view images. *Sci Rep.* 2021;11:4884.
87. Singh M, Venkataramani A. Rationing by Race [Internet]. National Bureau of Economic Research; 2022 [cited 2024 Oct 9]. Available from: <https://www.nber.org/papers/w30380>
88. Afshar M, Phillips A, Karnik N, Mueller J, To D, Gonzalez R, et al. Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. *J Am Med Inform Assoc.* 2019;26:254–61.
89. Zhou W, Prater LC, Goldstein EV, Mooney SJ. Identifying Rare Circumstances Preceding Female Firearm Suicides: Validating A Large Language Model Approach. *JMIR Mental Health.* 2023;10:e49359.
90. Sahai SY, Gurukar S, KhudaBukhsh WR, Parthasarathy S, Rempala GA. A machine learning model for nowcasting epidemic incidence. *Math Biosci.* 2022;343:108677.
91. Liao Q, Zhu M, Wu L, Pan X, Tang X, Wang Z. Deep Learning for Air Quality Forecasts: a Review. *Curr Pollution Rep.* 2020;6:399–409.
92. Peng Z, Zhang B, Wang D, Niu X, Sun J, Xu H, et al. Application of machine learning in atmospheric pollution research: A state-of-art review. *Sci Total Environ.* 2024;910:168588.
93. Mathis WS, Zhao S, Pratt N, Weleff J, De Paoli S. Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: How does it compare to traditional methods? *Comput Methods Programs Biomed.* 2024;255:108356.
94. Jalali MS, Akhavan A. Integrating AI language models in qualitative research: Replicating interview data analysis with ChatGPT. *Syst Dyn Rev.* 2024;40:e1772.
95. Qiao S, Fang X, Garrett C, Zhang R, Li X, Kang Y. Generative AI for Qualitative Analysis in a Maternal Health Study: Coding In-depth Interviews using Large Language Models (LLMs) [Internet]. *medRxiv*; 2024 [cited 2024 Nov 27]. p. 2024.09.16.24313707. Available from: <https://www.medrxiv.org/content/https://doi.org/10.1101/2024.09.16.24313707v1>
96. Bail CA. Can Generative AI improve social science? *Proc Natl Acad Sci.* 2024;121:e2314021121.
97. Lerner J. The Promise & Pitfalls of AI-Augmented Survey Research [Internet]. NORC at the University of Chicago. 2024 [cited 2024 Nov 27]. Available from: <https://www.norc.uchicago.edu/research/library/promise-pitfalls-ai-augmented-survey-research.html>
98. Long RA, Ballard S, Shah S, Bianchi O, Jones L, Koretsky MJ, et al. A new AI-assisted data standard accelerates interoperability in biomedical research [Internet]. *medRxiv*; 2024 [cited 2024 Nov 27]. p. 2024.10.17.24315618. Available from: <https://www.medrxiv.org/content/https://doi.org/10.1101/2024.10.17.24315618v1>
99. Tabatabaei Hosseini SA, Kazemzadeh R, Foster BJ, Arpali E, Süsal C. New Tools for Data Harmonization and Their Potential Applications in Organ Transplantation. *Transplantation.* 2024;108:2306.
100. DERNONCOURT F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. *J Am Med Inform Assoc.* 2017;24:596–606.
101. Tseng Y-J, Wang H-Y, Lin T-W, Lu J-J, Hsieh C-H, Liao C-T. Development of a Machine Learning Model for Survival Risk Stratification of Patients With Advanced Oral Cancer. *JAMA Network Open.* 2020;3:e2011768.
102. Corey KM, Kashyap S, Lorenzi E, Lagoo-Deenadayalan SA, Heller K, Whalen K, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): A retrospective, single-site study. *PLOS Med.* 2018;15:e1002701.

103. Maheshwari K, Ruetzler K, Saugel B. Perioperative intelligence: applications of artificial intelligence in perioperative medicine. *J Clin Monit Comput*. 2020;34:625–8.
104. Hyer JM, Ejaz A, Tsilimigras DI, Paredes AZ, Mehta R, Pawlik TM. Novel Machine Learning Approach to Identify Preoperative Risk Factors Associated With Super-Utilization of Medicare Expenditure Following Surgery. *JAMA Surg*. 2019;154:1014–21.
105. Tsoukalas A, Albertson T, Tagkopoulos I. From Data to Optimal Decision Making: A Data-Driven, Probabilistic Machine Learning Approach to Decision Support for Patients With Sepsis. *JMIR Med Inform*. 2015;3:e3445.
106. Giordano C, Brennan M, Mohamed B, Rashidi P, Modave F, Tighe P. Accessing Artificial Intelligence for Clinical Decision-Making. *Front Digit Health* [Internet]. 2021 [cited 2024 Aug 5];3. Available from: <https://www.frontiersin.org/journals/digital-health/articles/https://doi.org/10.3389/fdgh.2021.645232/full>
107. Magrabi F, Ammenwerth E, McNair JB, Keizer NFD, Hyppönen H, Nykänen P, et al. Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications. *Yearb Med Inform*. 2019;28:128–34.
108. Subcommittee on Social and Behavioral Sciences of the Committee on Science. Blueprint for the use of social and behavioral science to advance evidence-based policymaking [Internet]. National Science and Technology Council; 2024 May. Available from: <https://www.whitehouse.gov/wp-content/uploads/2024/05/Blueprint-for-the-Use-of-Social-and-Behavioral-Science-to-Advance-Evidence-Based-Policymaking.pdf>
109. Skeem JL, Lowenkamp CT. Risk, Race, and Recidivism: Predictive Bias and Disparate Impact. *Criminology*. 2016;54:680–712.
110. Pearl J. Causality. 2nd ed. Cambridge University Press; 2009. <https://doi.org/10.1017/CBO9780511803161>
111. Matthay EC, Hagan E, Gottlieb LM, Tan ML, Vlahov D, Adler NE, et al. Alternative causal inference methods in population health research: Evaluating tradeoffs and triangulating evidence. *SSM – Popul Health*. 2020;10:100526.
112. Textor J, Hardt J, Knüppel S. DAGitty: A Graphical Tool for Analyzing Causal Diagrams. *Epidemiology*. 2011;22:745.
113. Abadie A, Diamond A, Hainmueller J. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *J Am Stat Assoc*. 2010;105:493–505.
114. Ben-Michael E, Feller A, Rothstein J. The augmented synthetic control method. *J Am Stat Assoc*. 2021;116:1789–803.
115. Rudolph JE, Fox MP, Naimi AI. Simulation as a Tool for Teaching and Learning Epidemiologic Methods. *Am J Epidemiol*. 2021;190:900–7.
116. Van Der Laan MJ, Rose S. Targeted Learning: Causal Inference for Observational and Experimental Data [Internet]. New York, NY: Springer; 2011 [cited 2023 Jul 25]. Available from: <https://link.springer.com/https://doi.org/10.1007/978-1-4419-9782-1>
117. Gruber S, Logan RW, Jarrín I, Monge S, Hernán MA. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Stat Med*. 2015;34:106–17.
118. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data. *Epidemiology*. 2009;20:512.
119. Laan MJ van der, Polley EC, Hubbard AE. Super Learner. Statistical Applications in Genetics and Molecular Biology [Internet]. 2007 [cited 2024 Aug 8];6. Available from: <https://www.degruyter.com/document/doi/https://doi.org/10.2202/1544-6115.1309/html>
120. Luedtke A, Carone M, Simon N, Sofrygin O. Learning to learn from data: Using deep adversarial learning to construct optimal statistical procedures. *Sci Adv*. 2020;6:eaaw2140.
121. Luedtke A. Simplifying debiased inference via automatic differentiation and probabilistic programming [Internet]. arXiv; 2024 [cited 2025 Jan 22]. Available from: <http://arxiv.org/abs/2405.08675>
122. Rudolph KE, Gimbrone C, Díaz I. Helped into Harm: Mediation of a housing voucher intervention on mental health and substance use in boys. *Epidemiology*. 2021;32:336–46.
123. Rudolph KE, Laan MJ. Robust Estimation of Encouragement Design Intervention Effects Transported Across Sites. *J R Stat Soc Ser B: Stat Methodol*. 2017;79:1509–25.
124. Rudolph KE, Díaz I. Efficiently transporting causal direct and indirect effects to new populations under intermediate confounding and with multiple mediators. *Biostatistics*. 2022;23:789–806.
125. Chernozhukov V, Fernández-Val I, Luo Y. The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages. *Econometrica*. 2018;86:1911–38.
126. List JA, Shaikh AM, Xu Y. Multiple hypothesis testing in experimental economics. *Exp Econ*. 2019;22:773–93.
127. Baum A, Scarpa J, Bruzelius E, Tamler R, Basu S, Faghmous J. Targeting weight loss interventions to reduce cardiovascular complications of type 2 diabetes: a machine learning-based post-hoc analysis of heterogeneous treatment effects in the Look AHEAD trial. *Lancet Diabetes Endocrinol*. 2017;5:808–15.
128. Díaz I. Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*. 2020;21:353–8.
129. Hirshberg DA, Wager S. Augmented minimax linear estimation. *Ann Stat*. 2021;49:3206–27.
130. Naimi AI, Mishler AE, Kennedy EH. Challenges in Obtaining Valid Causal Effect Estimates With Machine Learning Algorithms. *Am J Epidemiol*. 2023;192:1536–44.
131. Kennedy EH. Semiparametric Doubly Robust Targeted Double Machine Learning: A Review. *Handbook of Statistical Methods for Precision Medicine* [Internet]. Chapman and Hall/CRC; 2024 [cited 2025 Jan 22]. p. 207–36. Available from: <https://www.taylorfrancis.com/chapters/edit/https://doi.org/10.1201/9781003216223-10/semiparametric-doubly-robust-targeted-double-machine-learning-review-edward-kennedy>
132. Hines O, Dukes O, Diaz-Ordaz K, Vansteelandt S. Demystifying Statistical Learning Based on Efficient Influence Functions. *Am Stat*. 2022;76:292–304.
133. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters. *Econom J*. 2018;21:C1–68.
134. Smith MJ, Phillips RV, Luque-Fernandez MA, Maringe C. Application of targeted maximum likelihood estimation in public health and epidemiological studies: a systematic review. *Annals of Epidemiology* [Internet]. 2023 [cited 2023 Jul 25]; Available from: <https://www.sciencedirect.com/science/article/pii/S1047279723001151>
135. Kim JJH, Um RS, Lee JWY, Ajilore O. Generative AI can fabricate advanced scientific visualizations: ethical implications and strategic mitigation framework. *AI Ethics* [Internet]. 2024 [cited 2024 Aug 14]; Available from: <https://doi.org/10.1007/s43681-024-00439-0>
136. Epstein Z, Hertzmann A. The Investigators of Human Creativity. Art and the science of generative AI. *Science*. 2023;380:1110–1.
137. Liang W, Zhang Y, Cao H, Wang B, Ding DY, Yang X, et al. Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis. *NEJM AI*. 2024;1:A10a2400196.
138. Anderson LB, Kanneganti D, Houk MB, Holm RH, Smith T. Generative AI as a Tool for Environmental Health Research Translation. *GeoHealth*. 2023;7:e2023GH000875.
139. Ganjavi C, Eppler MB, Pekcan A, Biedermann B, Abreu A, Collins GS, et al. Publishers' and journals' instructions to authors on

- use of generative artificial intelligence in academic and scientific publishing: bibliometric analysis. *BMJ*. 2024;384:e077192.
140. Flanagan A, Kendall-Taylor J, Bibbins-Domingo K. Guidance for Authors, Peer Reviewers, and Editors on Use of AI, Language Models, and Chatbots. *JAMA*. 2023;330:702–3.
 141. NOT-OD-23-149: The Use of Generative Artificial Intelligence Technologies is Prohibited for the NIH Peer Review Process [Internet]. [cited 2024 Nov 22]. Available from: <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-23-149.html>
 142. Lin Z. Beyond principlism: Practical strategies for ethical AI use in research practices [Internet]. arXiv; 2024 [cited 2024 Nov 22]. Available from: <http://arxiv.org/abs/2401.15284>
 143. Cooper AF, Grimmelmann J. The Files are in the Computer: Copyright, Memorization, and Generative AI [Internet]. arXiv; 2024 [cited 2024 Dec 30]. Available from: <http://arxiv.org/abs/2404.12590>
 144. Fernández-Loría C, Provost F. Causal Decision Making and Causal Effect Estimation Are Not the Same...and Why It Matters. *INFORMS J Data Sci*. 2022;1:4–16.
 145. Matthey EC, Hagan E, Gottlieb LM, Tan ML, Vlahov D, Adler N, et al. Powering population health research: Considerations for plausible and actionable effect sizes. *SSM - Population Health*. 2021;14: 100789.
 146. Zhang Z, Neill DB. Identifying Significant Predictive Bias in Classifiers [Internet]. arXiv; 2017 [cited 2024 Mar 19]. Available from: <http://arxiv.org/abs/1611.08292>
 147. Ravishankar P, Mo Q, McFowland E III, Neill DB. Provable detection of propagating sampling bias in prediction models. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2023;37(8):9562–9. <https://doi.org/10.1609/aaai.v37i8.26144>.
 148. Boxer KS, McFowland III E, Neill DB. Auditing predictive models for intersectional biases. arXiv preprint arXiv:2306.13064; 2023.
 149. Pamplin II JR, Wheeler-Martin K, Shroff R, Cerda M, Neill DB. Identifying heterogeneous treatment effects of the COVID-19 pandemic on non-fatal opioid overdose among New York State Medicaid enrollees. In: *Proceedings of the world congress of epidemiology*. 2024.
 150. Allen B, Neill DB, Schell RC, Ahern J, Hallowell BD, Krieger M, et al. Translating Predictive Analytics for Public Health Practice: A Case Study of Overdose Prevention in Rhode Island. *Am J Epidemiol*. 2023;192:1659–68.
 151. Allen B, Schell RC, Jent VA, Krieger M, Pratty C, Hallowell BD, et al. PROVIDENT: Development and Validation of a Machine Learning Model to Predict Neighborhood-level Overdose Risk in Rhode Island. *Epidemiology*. 2024;35:232–40.
 152. Athey S, Wager S. Policy Learning With Observational Data. *Econometrica*. 2021;89:133–61.
 153. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing Study Results: A Potential Outcomes Perspective. *Epidemiology*. 2017;28:553.
 154. Rudolph KE, Schmidt NM, Glymour MM, Crowder R, Galin J, Ahern J, et al. Composition or Context: Using Transportability to Understand Drivers of Site Differences in a Large-scale Housing Experiment. *Epidemiology*. 2018;29:199.
 155. Murphy SA. Optimal Dynamic Treatment Regimes. *J R Stat Soc Ser B: Stat Methodol*. 2003;65:331–55.
 156. Montoya LM, van der Laan MJ, Luedtke AR, Skeem JL, Coyle JR, Petersen ML. The optimal dynamic treatment rule superlearner: considerations, performance, and application to criminal justice interventions. *Int J Biostat*. 2023;19:217–38.
 157. Bracic A, Callier SL, Price WN. Exclusion cycles: Reinforcing disparities in medicine. *Science*. 2022;377:1158–60.
 158. Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, Huang P-S, et al. Ethical and social risks of harm from Language Models [Internet]. arXiv; 2021 [cited 2024 Oct 9]. Available from: <http://arxiv.org/abs/2112.04359>
 159. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366:447–53.
 160. Joyce K, Smith-Doerr L, Alegria S, Bell S, Cruz T, Hoffman SG, et al. Toward a Sociology of Artificial Intelligence: A Call for Research on Inequalities and Structural Change. *Socius*. 2021;7:2378023121999581.
 161. National Institute on Minority Health and Health Disparities. SchARE [Internet]. NIMHD. 2023 [cited 2024 Dec 11]. Available from: <https://www.nimhd.nih.gov/resources/schare/>
 162. Perlis RH. Five Big Ideas in AI and Health With JAMA+ AI Editor in Chief Roy Perlis [Internet]. JAMA+ AI. [cited 2024 Oct 9]. Available from: <https://edhub.ama-assn.org/jn-learning/video-player/18917394>
 163. Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, et al. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digit Med*. 2024;7:1–20.
 164. Office of Science Policy, National Institutes of Health. Artificial Intelligence in Research: Policy Considerations and Guidance [Internet]. Office of Science Policy. 2024 [cited 2024 Aug 14]. Available from: <https://osp.od.nih.gov/policies/artificial-intelligence/>
 165. Murphy K, Di Ruggiero E, Upshur R, Willison DJ, Malhotra N, Cai JC, et al. Artificial intelligence for good health: a scoping review of the ethics literature. *BMC Med Ethics*. 2021;22:14.
 166. Karimian G, Petelos E, Evers SMAA. The ethical issues of the application of artificial intelligence in healthcare: a systematic scoping review. *AI Ethics*. 2022;2:539–51.
 167. Oniani D, Hilsman J, Peng Y, Poropatich RK, Pamplin JC, Legault GL, et al. Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare. *npj Digit Med*. 2023;6:1–10.
 168. Lupton D, Butler E. Generative AI in Medicine and Public Health: An Overview and Position Paper on Directions for Social Research [Internet]. Rochester, NY; 2024 [cited 2024 Aug 14]. Available from: <https://papers.ssrn.com/abstract=4871308>
 169. Warraich HJ, Tazbaz T, Califf RM. FDA Perspective on the Regulation of Artificial Intelligence in Health Care and Biomedicine. *JAMA*. 2025;333:241–7.
 170. Nabi R, Benkeser D. Fair Risk Minimization under Causal Path-Specific Effect Constraints [Internet]. arXiv; 2024 [cited 2025 Jan 8]. Available from: <http://arxiv.org/abs/2408.01630>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.