Causal machine learning for predicting treatment outcomes

Stefan Feuerriegel[*,1,2], Dennis Frauen[1,2], Valentyn Melnychuk[1,2], Jonas Schweisthal[1,2], Konstantin Hess[1,2], Alicia Curth[3], Stefan Bauer[4,5], Niki Kilbertus[2,4,5], Isaac S. Kohane[6], and Mihaela van der Schaar[7,8]

[1]LMU Munich, Munich, Germany

[2]Munich Center for Machine Learning, Munich, Germany

[3]Department of Applied Mathematics & Theoretical Physics, University of Cambridge, Cambridge, United Kingdom

[4]School of Computation, Information and Technology, TU Munich, Munich Germany

[5]Helmholtz Munich, Munich, Germany

[6]Department of Biomedical Informatics, Harvard Medical School, Boston, USA

[7]Cambridge Centre for AI in Medicine, University of Cambridge, Cambridge, United Kingdom

[8]The Alan Turing Institute, London, United Kingdom

Abstract

Causal machine learning (ML) offers flexible, data-driven methods for predicting treatment outcomes. Here, we present how methods from causal ML can be used to understand the effectiveness of treatments, thereby supporting the assessment and safety of drugs. A key benefit of causal ML is that allows for estimating individualized treatment effects, as well as personalized predictions of potential patient outcomes under different treatments. This offers granular insights into when treatments are effective, so that decision-making in patient care can be personalized to individual patient profiles. We further discuss how causal ML can be used in combination with both clinical trial data as well as real-world data such as clinical registries and electronic health records. We finally provide recommendations for the reliable use of causal ML in medicine.

First published in Nature Medicine, 30, 958–968 (2024) by Springer Nature. Link: https://doi.org/10.1038/s41591-024-02902-1

[*]Corresponding author: feuerriegel@lmu.de

Main

Assessing the effectiveness of treatments is crucial to ensure patient safety and personalize patient care. Recent innovations in machine learning (ML) offer new, data-driven methods to estimate treatment effects from data. This branch in ML is commonly referred to as causal ML as it aims to predict a causal quantity, namely, the patient outcomes due to treatment [1]. Causal ML can be used in order to estimate treatment effects from both experimental data obtained through randomized controlled trials (RCTs) and observational data obtained from clinical registries, electronic health records, and other real-world data (RWD) sources to generate clinical evidence. A key strength of causal ML is that it allows to estimate individualized treatment effects, as well as to make personalized predictions of potential patient outcomes under different treatments. This offers a granular understanding of when treatments are effective or harmful, so that decision-making in patient care can be personalized to individual patient profiles.

Box 1. Glossary of common terms in causal ML

- Causal graph: A graphical representation of the causal relationships between variables, typically using directed acyclic graphs to depict causal paths.
- Causal ML: A branch of machine learning that aims at the estimation of causal quantities (e.g., average treatment effect, conditional average treatment effect) or at predicting potential outcomes. Here, "causal" implies that the target is a causal quantity when certain assumptions about the data-generating mechanism are satisfied. For alternative definitions and use cases of causal ML, see [1].
- Confounder: A variable that influences both the treatment assignment and the outcome.
- Consistency: The potential outcome is equal to the observed patient outcome under the selected treatment, which implies that the outcomes are clearly defined.
- Counterfactual outcome: The unobservable patient outcome that would have occurred, had a patient received a different treatment.
- Factual outcome: The observed patient outcome that occurred for the observed treatment.
- Identifiability: A statistical concept referring to the ability of whether causal quantities such as treatment effects can be uniquely inferred from the observed data.
- Positivity: Each patient has a bigger-than-zero probability of receiving/not receiving a treatment. This is also called overlap assumption.
- Potential outcome: The hypothetical patient outcome that would be observed if a certain treatment was administered.
- Propensity score: The propensity score is the probability of receiving the treatment given the observed specific patient characteristics.
- Stable unit treatment value (SUTVA): The outcome for any patient does not depend on the treatment assignment of other patients, and there is no variation in the effect of the treatment across different settings or populations.
- Unconfoundedness: Given observed covariates, the assignment to treatment is independent of the potential outcomes. This is the case e.g., when there are no unobserved confounders, that is, variables influencing both the treatment and the outcome. The assumption is also called ignorability.

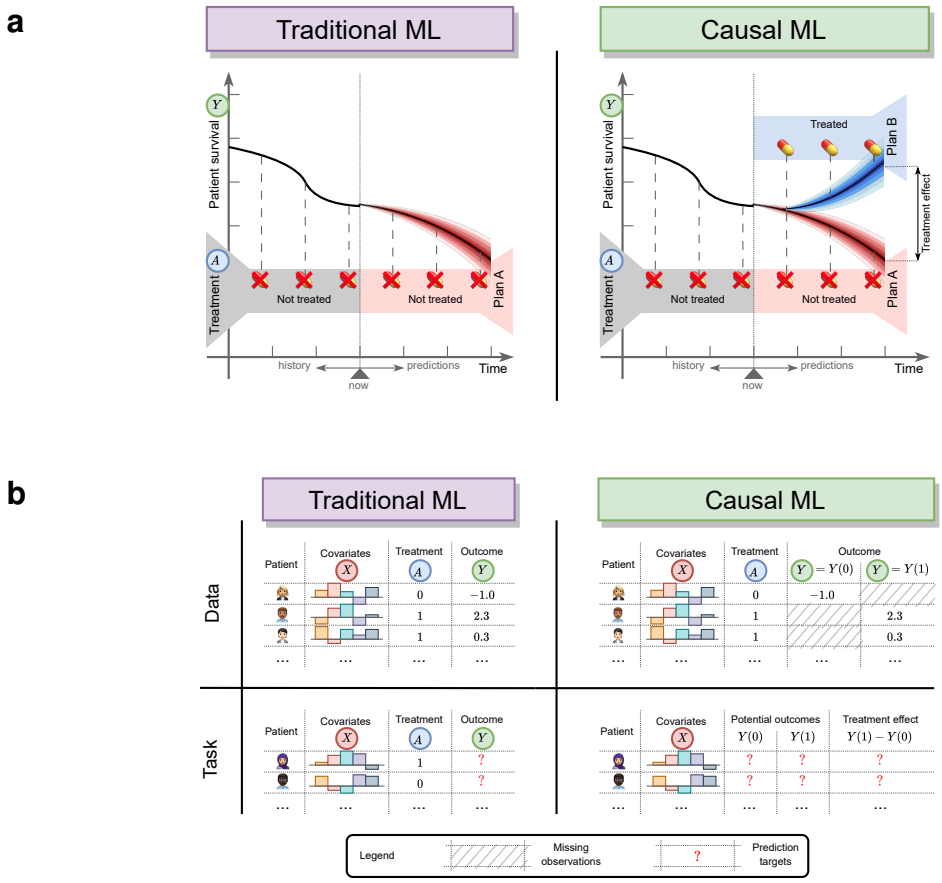

Figure 1: **Causal ML for predicting treatment outcomes.** **a**, Different from traditional ML, causal ML aims at (i) estimating the treatment effect or (ii) predicting the patient outcomes themselves due to treatments. **b**, Causal ML is challenging due to the fundamental problem of causal inference in that not all potential outcomes can be observed and are thus missing in the data. **c**, Treatment effect heterogeneity.

Causal ML in medicine

Comparison to traditional ML

Causal ML for estimating treatment effects is different from traditional predictive ML (see Box 1 for a glossary of terms). Intuitively, traditional ML aims at predicting outcomes [2], while causal ML quantifies changes in outcomes due to treatment, so that treatment effects can be estimated (see Fig. 1a). For example, a typical use case for traditional ML is risk scoring, such as predicting the probability of a diabetes onset to understand which patients are at high risk but without saying what the best treatment plan is [3–7]. In contrast, causal ML aims to answer "what if" questions. For example, causal ML could answer the question of how the risk of a diabetes onset will *change* if the patient receives an antidiabetic drug [8–10], so that decisions can be made whether to administer an antidiabetic drug. Causal ML can also be used to predict the potential patient outcomes in response to different treatments. For example, in oncology, causal ML could make individualized predictions of survival under different treatment plans, which can then help medical practitioners in choosing a treatment plan that promises the largest chance of survival [11].

Estimating treatment effects from data requires custom methods. The reason is that treatment effects for individual patients are not observable. This is due to the so-called fundamental problem of causal inference [12, 13]: one can only observe the (factual) patient outcome under the given treatment, but one never observes the (counterfactual) patient outcome under a different, hypothetical treatment (see Fig. 1b). Therefore, the estimation of treatment effects or other causal quantities, which are based on such unobserved outcomes, poses additional challenges not present in traditional, predictive ML.

To then obtain a causal quantity that can be estimated, certain assumptions on the causal structure of the problem must be made. In particular, one often needs to assume that there is no unmeasured confounding, that is, there are no unobserved factors that drive both treatments and patient

outcomes. If unmeasured confounding is present, the estimated treatment effects may suffer from confounding bias and, as a result, can even have a wrong sign [14]. Additionally, to estimate treatment effects, one needs to account for the dependence structure between treatment, outcomes, and patient characteristics by modeling the underlying causal relationships. One reason is that intervening on the treatment variable could also affect other patient characteristics. As an example, consider a patient with a certain body mass index for whom the diabetes risk should be predicted and where the doctor recommends stopping smoking. Literature from traditional ML would suggest using both the body mass index and smoking behavior to predict the diabetes risk under smoking vs. no smoking; however, this approach would ignore that stopping smoking would also change a patient's body mass index. Instead, ML needs to be embedded in a causal framework.

Benefits

Methods for estimating treatment effects have a long tradition in the statistics literature (e.g., [15–18]). Causal ML builds upon the same problem setup but makes changes to the estimation strategy. Hence, the core improvement from using causal ML is generally not the types of questions that can be asked, but how these questions can be answered. As such, causal ML can have benefits over alternative methods from the statistics literature (see Box 2). First, methods from classical statistics often assume knowledge about the parametric form of the association between patient characteristics and outcomes, such as linear dependencies. However, such knowledge is often not available or unrealistic, especially for high-dimensional datasets such as electronic health records, and this can easily lead to models that are misspecified. In contrast, causal ML typically allows for less rigid models, which helps in capturing complex disease dynamics as well as human pathophysiology and pharmacology. Still, there is a trade-off as causal ML typically requires larger sample sizes. For example, parsimonious models such as linear regressions are often favorable for settings with small sample sizes, while more flexible, non-linear models are only effective for large sample sizes.

Box 2. Comparison of causal ML vs. traditional statistics

Due to the importance of treatment effect estimation across many application areas, methods for treatment effect estimation have been developed in different disciplines, including statistics, biostatistics, econometrics, and machine learning (e.g., [19–26]). However, there is no 'dichotomy' as many concepts are shared across disciplines. For example, many state-of-the-art methods for estimating treatment effects are model-agnostic in that they can be used in combination with both arbitrary models from classical statistics and also more modern machine learning models [19–21].

Eventually, the choice of whether to rely on a more classical statistical model or a more modern machine learning method presents a trade-off that depends on the underlying settings. For example, simple models (e.g., linear regression or other parametric models) are often preferred for small sample sizes. For large sample sizes, more complex, non-linear models can be used to capture heterogeneity in the treatment effect. Notwithstanding, the ability to handle non-linear relationships and treatment effect heterogeneity is not unique to causal ML but can, in principle, also rely on classical statistical models that allow incorporating prespecificed non-linearities. Therefore, causal ML may have advantages when the underlying data-generating process is complex and when prior knowledge is limited.

In medicine, causal ML offers several opportunities for estimating treatment effects from data, which eventually help in greater personalization of care. First, at the patient level, causal ML can handle high-dimensional and unstructured data with patient covariates and thus estimate treatment effects from multi-modal datasets containing images, text, or time series, as well as genetic data. For example, one could estimate treatment effects from computed tomography (CT) scans or entire electronic health records. Second, at the outcome level, causal ML can help make personalized estimates of treatment effects for subpopulations or even predict outcomes for individual patients

[27]. For example, individual differences in drug metabolism can lead to serious side effects for drugs in some patients but can be lifesaving in others [28], so causal ML could learn such differences and thus help in designing personalized treatment strategies. Third, at the treatment level, causal ML can be effective for estimating heterogeneity in treatment effects across patients in a data-driven manner to better understand where treatments are effective (see Fig. 1c).

Workflow

The workflow in Fig. 2 outlines the different steps necessary to predict treatment outcomes with causal ML. The workflow [29,30] should help researchers in clearly defining the research question and then guide the choice of the causal quantity of interest, the causal model, the causal ML method, and the robustness checks to validate the reliability of the estimates.

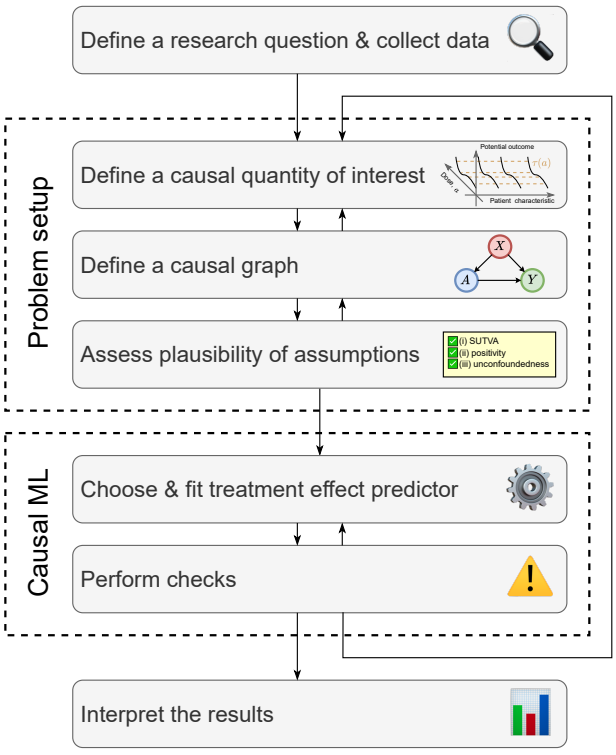

Figure 2: **Workflow for causal ML in medicine.**

Problem setup

To estimate the effectiveness of treatments, information about the following variables is necessary [13]: the treatment of interest; the observed patient outcome; and patient characteristics such as age, gender, and the medical history. The patient characteristics are commonly also called covariates. For example, in cancer care, one could use electronic patient records with information about the type of chemotherapy (the treatment), the size of a cancer tumor (the outcome), and the previous medical history (the covariates). In the standard setting [13], the variables can influence each other as shown by the example causal graph in Fig. 3a. To make causal quantities identifiable, we later need to assume knowledge about the causal graph.

Information about the above variables can come from either observational or experimental data. In observational data, the treatment assignment is unknown and not fully randomized. This is the case in RWD such as clinical registries and electronic health records. Here, the treatment assignment followed some, typically unknown procedure depending on the patient characteristics. For example, patients with a more severe illness are likely to get a more aggressive form of treatment, implying that the patient characteristics differ across treatment groups. This is unlike RCTs, where treatments are randomized, which is also referred to as experimental data. As a result, the patient characteristics are similar across treatments. This is captured by the propensity score, which is the probability of receiving a treatment given the patient covariates [15]. In RCTs, the propensity score is known (e.g., the propensity score is 50% in completely randomized trials with two treatment arms of equal size). In contrast, the treatment assignment in RWD is unknown but can be estimated to account for differences in the patient populations of those who have received treatment and those who have not.

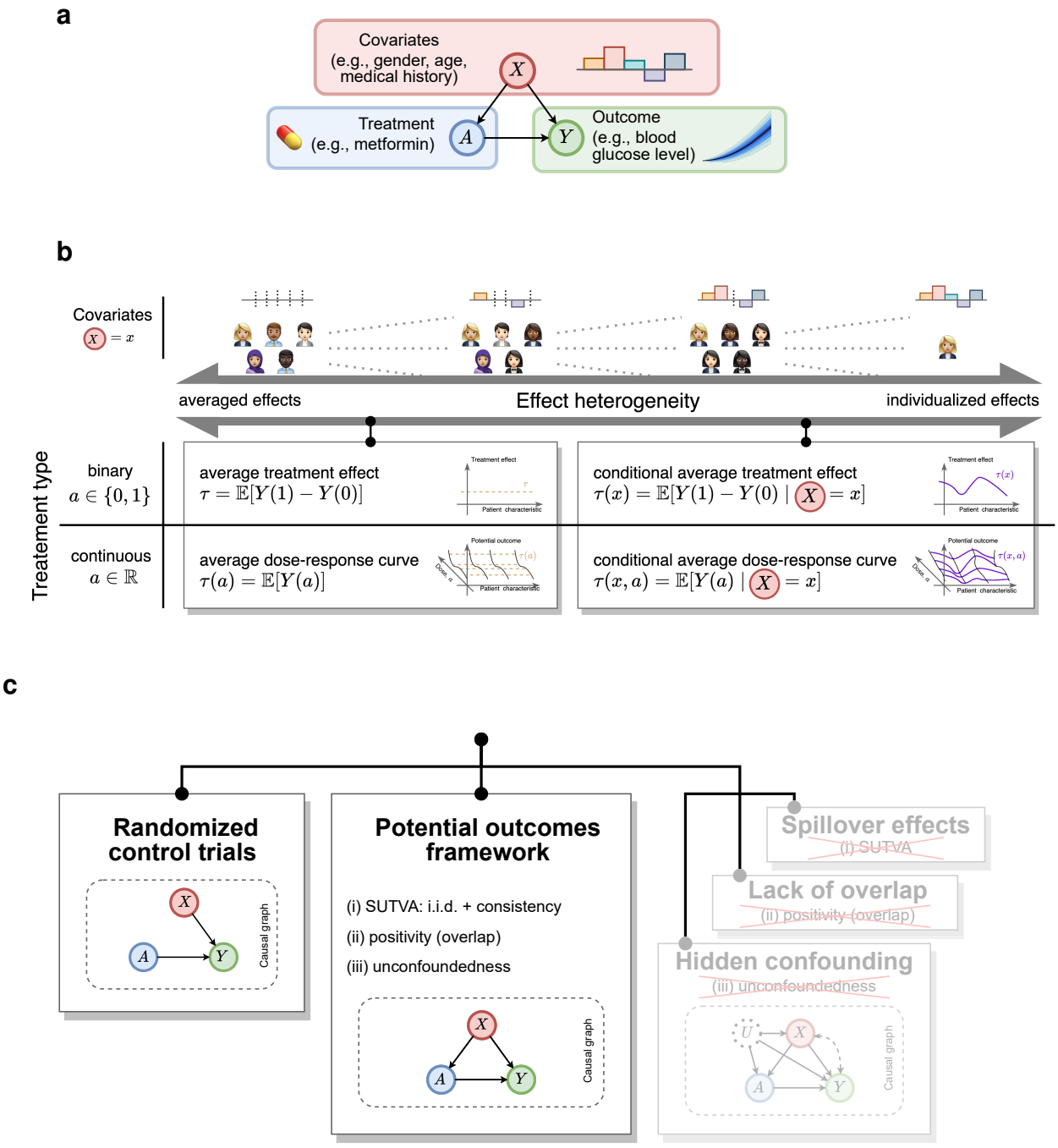

Figure 3: **Formalizing tasks for causal ML. a**, A causal graph must be assumed such as the example here. The arrows ($\rightarrow$) indicate the causal relationships between different variables. Note that the causal graph allows for possible unobserved variables (not confounders) that are correlated with treatment and confounders, or correlated with confounders and outcome. **b**, The research question defines what causal quantity is of interest, that is, the so-called estimand. The estimand can vary by the effect heterogeneity (average vs. individualized) and treatment type (binary vs. continuous). Here, $Y(a)$ is the potential outcome for treatment a. **c**, The research question determines the problem setup, which comes with different assumptions that must be made to ensure identifiability and thus reliable inferences.

Types of treatment effects

Outcomes of treatments are commonly formalized based on the potential outcomes framework [15]. The framework conceptualizes so-called potential outcomes, which are the patient outcomes that would hypothetically be observed if a certain treatment was administered. Then, depending on the practical applications, different causal quantities can be of interest. These include treatment effects, which quantify the expected difference of two potential outcomes under different treatments. Common choices of treatment effects can be loosely grouped along two dimensions (see Fig. 3b): (i) the degree of effect heterogeneity and (ii) the treatment type. By choosing a specific treatment effect of interest, one defines the so-called estimand, that is, the causal quantity that should be estimated by the causal ML method.

(i) *Degree of effect heterogeneity*: Traditionally, average treatment effects (ATEs) are widely used in clinical trials [15]. ATEs measure effects at the level of the study population. By comparing the average patient outcome for those receiving the treatment versus those who do not (control group), ATEs help in understanding how effective a treatment is, on average, across a specific patient cohort [19]. This is important, for example, when analyzing the comparative effectiveness of a new drug compared to the standard of care, or when assessing the overall effectiveness or safety of a new drug. However, ATEs can not offer granular insights into whether patients with specific covariates may particularly benefit from a treatment. Nevertheless, such heterogeneity in treatment effects can be of high interest in practice (see Fig. 1c). For this, one typically estimates the conditional average treatment effect (CATE), which is the effect of a treatment for a particular subgroup of patients defined by the covariates. Understanding the heterogeneity in treatment effects informs about subgroups where treatments are not effective or even harmful, which is relevant for individualizing treatment recommendations to specific patients.

(ii) *Treatment type*: Binary treatments refer to a type of treatment variable that is dichotomous and thus has only two categories, for example, when answering questions of whether to treat or

not to treat. In contrast, discrete as well as continuous treatments refer to a type of treatment variable that can take on a range of values rather than being limited to two (or a few) categories. Continuous treatment variables are commonly present in situations where the intensity, dosage, or level of exposure to treatment can be flexibly chosen [31]. For example, in radiation therapy, the dose of radiation is often chosen from a fairly wide spectrum that depends on the cancer type and other patient characteristics [32]. For continuous treatments, the treatment effectiveness is often also summarized by dose-response curves.

Besides the above, some applications in medicine are also interested in the potential outcomes themselves, rather than in treatment effects. While understanding the treatment effectiveness is often important for the assessment and safety of drugs, the potential outcomes can support decision-making in routine care by helping clinical professionals reason about what outcome to expect under different treatment options. This may be seen as a 'risk under intervention' estimand and requires a careful modeling strategy [33]. For example, while the treatment effect may say that a drug can improve the 5-year mortality by 5 percentage points, the predicted outcomes inform that the mortality is 20% with treatment and 15% without, so that not only the relative gain but the future health state can be considered. However, in practice, the estimation of ATE and CATE is often an easier task than estimating potential outcomes [24] and, hence, is preferred when it is sufficient for decision-making.

Assumptions for identifiability

The estimation of treatment effects involves counterfactual outcomes, which are not observable. Therefore, formal assumptions must be made about the data-generating process to ensure the identifiability of treatment effects from data [13]. Intuitively, identifiability is a theoretical concept that refers to whether causal quantities such as treatment effects can be uniquely inferred from data. Ensuring identifiability is necessary since, otherwise, it is impossible in practice to estimate a treatment effect without bias even with infinite data [13].

RCTs ensure the identifiability of treatment effects through fully randomized treatment assignment. However, treatment assignment in RWD is unknown and not fully randomized, namely, it depends on other covariates, so that formal assumptions must be made [15]. The exact set of assumptions depends on which type of treatment effect is chosen. For the treatment effects discussed above, the following three assumptions (see Fig. 3c) are standard [15, 34]: (i) Stable unit treatment value assumption (SUTVA) implies that the RWD are independent and identically distributed and that consistency holds (i.e., the potential outcome coincides with the observed outcome for a given treatment). This assumption implies that there is no interference between patients in which treating one patient influences the outcomes for another patient in the study population (e.g., due to spillover or peer effects). (ii) Positivity (also called overlap) requires a non-zero probability of receiving a treatment. Positivity implies that, for each possible combination of patient characteristics, we can observe both treated and untreated patients. (iii) Unconfoundedness (also called ignorability) states that, given observed covariates, the treatment assignment is independent of the potential outcomes. In particular, this is satisfied if the patient covariates include all possible confounders, i.e., variables that influence both the treatment and the outcome. For example, unconfoundedness may be violated if patients with certain sociodemographics such as race or high income tend to have better access to treatments and better adherence [35] and where the reason is not captured in the data. In principle, unconfoundedness can be addressed by capturing all relevant factors behind driving treatment assignment in RWD [36], yet it is generally challenging to validate this in practice. If confounders are not observed or not modeled (or even not known), then not only might the magnitude of the estimated treatment effect be biased but it might even have the wrong sign [14].

Importantly, assumptions such as those above are required for consistently estimating treatment effects from data, regardless of which method is used. A natural challenge is further that assessing the plausibility of the assumptions is often difficult. Later, we discuss potential strategies to check the credibility of whether the assumptions hold. Notwithstanding, there are also problem setups with other designs. For example, some problem setups allow for relaxations of the SUTVA as-

sumption (e.g., by allowing for spillover effects) [37,38]. There also exist alternatives to assuming unconfoundedness in specific settings such as through the use of instrumental variables [39, 40]. Finally, there are problem setups that are not static but time-varying, so that a sequence of treatment decisions is made over time [41–50]. Even other works focus how to effectively combine both observational and experimental data [51–53].

Methods

There are different causal ML methods, which vary based on which causal quantity of interest and which causal model is addressed. For example, a large body of literature focuses on causal ML for ATE [19, 20, 54–57]. For CATE estimation with binary treatments, there are two broader categories of methods (see Box 3). On the one hand, so-called meta-learners [22] are model-agnostic methods for CATE estimation that can be used for treatment effect estimation in combination with an arbitrary ML model of choice (e.g., a decision tree, a neural network [23]). On the other hand, model-specific methods make adjustments to existing ML models to address statistical challenges arising in treatment effect estimation and, therefore, to improve performance. Here, prominent examples are the causal tree [58] and the causal forest [59, 60], which adapt the decision tree and random forest, respectively, for treatment effect estimation. Even others adapt representation learning to leverage neural networks for treatment effect estimation [61, 62]. In contrast, different methods are needed for continuous treatments such as in settings where the intensity, dosage, or level of exposure to a treatment can be flexibly chosen [21, 21, 31, 63–69]. The reason is that this poses a challenging task as the number of treatment values is infinite and not every value is observed in the data.

Existing causal ML methods often focus on point estimates. This can be a serious limitation in medical applications [70], where uncertainty estimates such as standard errors or confidence intervals are crucial for reliable decision-making [71]. However, there is also some progress. For example, for CATE estimation, the causal forest [59,60] is a method that offers rigorous uncertainty

estimates. In addition, several other strategies have been developed recently, such as Bayesian methods [72] and conformal prediction [73], but still more research is needed.

Box 3. Meta-learners in causal ML

Meta-learners [22] are model-agnostic methods for CATE estimation that can be used for treatment effect estimation in combination with an arbitrary ML model of choice (e.g., a decision tree, a neural network [23]). There are different ways in which such meta-learners can leverage the data in a supervised learning setting.

Plug-in learners: One way is to train a single ML model that predicts the patient outcome but where the treatment is added as a separate variable to the covariates (called S-learner [22]). Another way is to train two separate ML models for each treatment (called T-learner [22]). Here, one ML model is trained for predicting patient outcomes in the treatment group and one ML model for the control group. After having computed the ML model(s), one simply uses the estimated treated and control outcome to "plug them into" the formula for computing the treatment effect.

Two-step learners: An alternative approach is to target the CATE, which can lead to faster convergence [24]. However, Because the difference between factual and counterfactual outcomes is never observed in data, so-called pseudo-outcomes are used as surrogates, which have the same expected value as the CATE. Prominent examples are the so-called DR-learner [24] and the so-called R-learner [25], which come with certain robustness guarantees [21, 25, 26].

The above meta-learners have different advantages and disadvantages. Unfortunately, there are no clear rules for choosing meta-learners but only high-level recommendations [23, 74].

Evaluation

Arguably, the best way to evaluate models is to access randomized data. While this does not allow to assess treatment effects of individual patients, it still helps during model selection, so that models are favored with the best performance in terms of average or heterogeneous treatment effects. In contrast, benchmarking for the purpose of model selection is challenging, as both counterfactuals and ground-truth values of treatment effects are unknown [75–77]. As a remedy, two strategies are common. (1) A simple heuristic is to compare methods from causal ML based only on the performance in predicting factual outcomes, whereby the performance in predicting counterfactual outcomes is ignored. This heuristic may give some insights into whether the underlying disease mechanisms in the data are captured. Yet, it has a major limitation in that the key causal quantity of interest – i.e., the treatment effect – is not evaluated. (2) Another heuristic is to use pseudo-outcomes [78]. Here, a pseudo-outcome is first estimated using a secondary, independent model to approximate the unknown counterfactual outcome, and, then, the pseudo-outcome is used to benchmark the estimated CATE. However, such a heuristic depends on the performance of the secondary model for pseudo-outcomes and tends to favor certain methods [78]. Still, the two strategies are heuristics and thus have inherent limitations.

Technical recommendations

Checking the plausibility of assumptions

Assessing the plausibility of the underlying assumptions is crucial for the validity of treatment effect estimates, yet also challenging. For the consistency assumption, one should assert that the treatment of one patient does not affect the outcome of another based on domain knowledge. For the positivity assumption, one typically plots the propensity scores to check whether the propensity scores are not too small or too large, since, otherwise, there may not be enough support in the data

for reliable inferences [79]. Another strategy is to rely upon methods for uncertainty quantification as some treatments may be given rarely to certain patient cohorts, implying that there may be limited support in the data for making inferences in these patient cohorts and, therefore, a large uncertainty [80]. If the positivity assumption is violated, a strategy is to exclude certain subgroups from the analysis as no reliable inferences for them can be made [79, 81].

Validating the unconfoundedness assumption is especially challenging for RWD. The best way to avoid violations of the unconfoundedness assumption is to consult domain knowledge to ensure that all relevant factors behind treatment assignment are captured in RWD [36]. An alternative is to adopt an instrumental variable approach [39, 40] but valid instruments are often rare in medical applications and, again, the validity of instruments cannot be tested. If unobserved confounders cannot be ruled out, conducting a causal sensitivity analysis can be helpful to assess how robust the results are to potential unobserved confounding. Causal sensitivity analysis dates back to a study from 1959 showing that unobserved confounders cannot explain away the causal effect of smoking on cancer [82]. Causal sensitivity analysis computes bounds on the causal effect of interest under some restriction on the amount of confounding, thus implying that a treatment effect cannot be explained away. Then, restrictions on the amount of confounding are based on domain expertise, typically by making comparisons to known, important causes that act as baselines (e.g., risk factors such as age). Recently, a series of causal ML methods have been proposed that provide sharp bounds [83–87]. However, causal sensitivity analysis still requires that there is sufficient knowledge of human pathophysiology and pharmacology about important disease causes, which may not always be the case in observational studies [14].

In addition, there are so-called refutation methods to validate the robustness of the treatment effect estimates against explicit violations of the different assumptions [88]. Common refutation methods are, for example, adding a random variable to check if the treatment effect estimates remain consistent as such variable should not affect the estimates, or replacing the actual treatment variable with a random variable to check if the estimated treatment effect goes to zero. Further,

one could perform simulations where the outcome is replaced through semi-synthetic data to check if the treatment effect is correctly estimated under the new data-generating mechanism for the simulated outcomes. Altogether, the choice of which refutation method to use for validating the used methods highly depends on the specific problem setting and should be carefully chosen and implemented. Still, even when the refutation methods yield a positive result, this is no guarantee that the assumptions are satisfied.

Notwithstanding, robustness checks that are best practice in ML are still essential (e.g., to mitigate the risk of bias [89]), especially as the results in treatment effect estimation may heavily depend on both the data and the model choice.

Reporting

Findings should be interpreted with great care. In particular, the assumptions, the rationale for the chosen causal ML method, and the robustness checks should be clearly stated. If possible, the estimated treatment effects from RWD should be compared against those from RCTs. This can help in validating the reliability of the causal ML methods but may also reveal differences between clinical trials and routine care (e.g., due to different patient cohorts or different levels of adherence). The reliability of the estimated treatment effects also depends on the quality and representativeness of the underlying data. Furthermore, analyses through causal ML involve multiple hypotheses testing and, therefore, are at risk of false positives. Similarly, due to the retrospective nature of such analyses, another risk is selective reporting of positive results. To mitigate such risks, preregistered protocols for analysis are highly recommended [90, 91]. Finally, when causal ML is used together with RWD, the limitations of making causal conclusions should be openly acknowledged, and, if possible, RCTs should be considered for validation.

Clinical translation

By estimating treatment effects from medical data, causal ML offers significant potential to personalize treatment strategies and improve patient health. Still, there is a long way to go. A key focus for future research must be on bridging the gap between ML research and direct benefits for patients in clinical practice.

Use cases

Causal ML can help in generating new clinical evidence. For RCTs, causal ML may determine specific patient cohorts within the population that might respond positively (or negatively) to a particular treatment. For example, the treatment effect of antidepressant drugs over a placebo varies substantially and tends to increase with baseline severity of the depression [92]. However, RCTs typically compare patient outcomes across two (or more) treatment arms, which would return the average treatment effect at the population level, and the use of causal ML may help to define inclusion criteria for clinical trials or to identify predictive biomarkers (e.g., certain genetic mutations in a tumor).

Furthermore, causal ML may offer flexible, data-driven methods to analyze treatment effect heterogeneity in RWD such as clinical registries and electronic health records. This is relevant as RCTs can be subject to limitations [93], such as high costs or that treatment randomization can be unethical for vulnerable populations (e.g., pregnant women) [94]. RWD together with causal ML could allow to estimate heterogeneous treatment effects for vulnerable groups, rare diseases, long-term outcomes, and uncommon side effects that are often not sufficiently captured by traditional RCTs. For example, as randomizing hospitalizations is typically not possible, one study used causal ML to estimate the effect of hospitalizations on suicide risk from RWD [95]. Likewise, patient populations in RCTs are often not representative of the broader population [96], but which one can account for through causal ML [97] in order to better understand the post-approval efficacy

of treatments. However, while the potential of RWD has been widely recognized [93, 98], many methodological questions are still unanswered, and causal ML may thus help in translating data into clinical evidence.

Eventually, the choice of the specific estimand depends on the setting where causal ML is used. For regulatory bodies, it may be relevant to assess the overall net benefit for patients at large, for example, when comparing a new drug against the standard of care. This would require the estimation of the average treatment effect. To ensure patient safety, regulatory bodies could also assess how the treatment effect varies across different subpopulations, which would involve the conditional average treatment effect. Likewise, the conditional average treatment effect may help to identify subpopulations that are particularly responsive to a treatment (e.g., for hypothesis generation) or that would benefit from newly developed drugs, thereby contributing to an accelerated drug development. When causal ML is integrated into clinical decision support systems in routine care, clinical professionals may want to make personalized predictions of how a patient's health state changes under different treatment options. This would require methods for CATE estimation or even for predicting potential patient outcomes.

Challenges and future directions

Several challenges in the clinical translation of causal ML are at the technical level. First, both estimating heterogeneous treatment effects as well as predicting patient outcomes are naturally difficult. In practice, this often requires both strong predictors of treatment effects and large sample sizes. While the former depends on the human pathophysiology and pharmacology in the specific disease setup, the latter may improve over time with an increasing prevalence of electronic health records. Another challenge is that uncertainty quantification for many causal ML methods is lacking. However, uncertainty quantification is crucial for reliable decision-making and thus for building clinical evidence [71]. For example, point estimates might indicate substantial effect heterogeneity, especially in settings with limited data, while, in fact, there is little heterogeneity

but simply large (aleatoric) uncertainty as the outcomes are difficult to predict. Hence, causal ML methods that only provide point estimates without conveying the appropriate uncertainty in the predictions may lead to potentially misleading or inappropriate conclusions. Finally, many causal ML methods are only implemented in specialized software libraries. Hence, comprehensive software tools are needed that improve reliability and ease of use and account for needs in medicine (e.g., rigorous uncertainty quantification).

The development of standardized protocols, ethical guidelines, and regulatory frameworks for causal ML applications will be essential in ensuring safe and effective treatment decisions. For example, tailored checklists based on consensus statements will need to be developed. While there are checklists for traditional, predictive ML [99] and for generating real-world evidence [91, 100], future research is needed that adapts such checklists to account for the needs of causal ML in medicine. Likewise, customized review processes will need to be developed, which define how evidence generated through the causal ML method must undergo regulatory review for approval.

So far, research in causal ML has primarily evaluated the performance of different methods through simulations (e.g., [41, 43, 44, 46, 48, 50]). However, simulations involve (semi-)synthetic datasets that do not fully capture the nuances of real-world disease dynamics. Hence, demonstrating the clinical insights generated through a cautious use of innovative causal ML methods can provide an important first step. This will help in understanding the strengths and limitations of causal ML in a medical context, especially in comparison to established clinical trials. For this, settings may appear especially suited where clear guidelines are missing, so that clinical decision support through causal ML can provide additional input by augmenting the decision-making of clinical professionals. Eventually, tools based on causal ML may be integrated into routine care through clinical decision support systems. Such systems may directly predict individual patient outcomes for different treatment options and thereby support the decision-making of clinical professionals.

Conclusion

Causal ML offers significant potential to draw novel conclusions about the efficacy and safety of treatments. In particular, causal ML offers abundant opportunities to personalize treatment strategies and thus to improve patient health. However, several challenges arise in practice. One challenge is to ensure the reliability and robustness of causal ML methods. Another challenge is to overcome barriers in the clinical translation. Here, proof-of-concept studies and a cautious use in practice can be an important first step.

Acknowledgments

SF acknowledges funding via the Swiss National Science Foundation (SNSF), Grant 186932.

Author contributions

All authors contributed to conceptualization, manuscript writing, and approved the manuscript.

Competing interests

The authors declare no competing interests.

References

[1] Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J. & Silva, R. Causal machine learning: A survey and open problems. *arXiv:2206.15475* (2022).

[2] Esteva, A. *et al.* A guide to deep learning in healthcare. *Nature Medicine* **25**, 24–29 (2019).

[3] Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A. & Stiglic, G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports* **10**, 11981 (2020).

[4] Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. & Van der Schaar, M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLOS ONE* **14**, e0213653 (2019).

[5] Cahn, A. *et al.* Prediction of progression from pre-diabetes to diabetes: Development and validation of a machine learning model. *Diabetes/Metabolism Research and Reviews* **36**, e3252 (2020).

[6] Zueger, T. *et al.* Machine learning for predicting the risk of transition from prediabetes to diabetes. *Diabetes Technology & Therapeutics* **24**, 842–847 (2022).

[7] Krittanawong, C. *et al.* Machine learning prediction in cardiovascular diseases: A meta-analysis. *Scientific Reports* **10**, 16057 (2020).

[8] Xie, Y. *et al.* Comparative effectiveness of SGLT2 inhibitors, GLP-1 receptor agonists, DPP-4 inhibitors, and sulfonylureas on risk of major adverse cardiovascular events: Emulation of a randomised target trial using electronic health records. *The Lancet Diabetes & Endocrinology* **11**, 644–656 (2023).

[9] Deng, Y. *et al.* Comparative effectiveness of second line glucose lowering drug treatments using real world data: emulation of a target trial. *BMJ Medicine* **2**, e000419 (2023).

[10] Kalia, S. *et al.* Emulating a target trial using primary-care electronic health records: Sodium-glucose cotransporter 2 inhibitor medications and hemoglobin a1c. *American Journal of Epidemiology* **192**, 782–789 (2023).

[11] Petito, L. C. *et al.* Estimates of overall survival in patients with cancer receiving different treatment regimens: Emulating hypothetical target trials in the Surveillance, Epidemiology, and End Results (SEER)–Medicare linked database. *JAMA Network Open* **3**, e200452 (2020).

[12] Holland, P. W. Statistics and causal inference. *Journal of the American Statistical Association* **81**, 945–960 (1986).

[13] Pearl, J. *Causality: Models, Reasoning, and Inference* (Cambridge University Press, 2009).

[14] Hemkens, L. G. *et al.* Interpretation of epidemiologic studies very often lacked adequate consideration of confounding. *Journal of Clinical Epidemiology* **93**, 94–102 (2018).

[15] Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701 (1974).

[16] Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* **100**, 322–331 (2005).

[17] Robins, J. M. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics: Theory and Methods* **23**, 2379–2412 (1994).

[18] Robins, J. M. Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association on Bayesian Statistical Science* 6–10 (1999).

[19] van der Laan, M. J. & Rubin, D. Targeted maximum likelihood learning. *The International Journal of Biostatistics* **2**, Article 11 (2006).

[20] van der Laan, M. J. & Rose, S. *Targeted learning: Causal inference for observational and experimental data* (Springer, New York, NY, 2011).

[21] Foster, D. J. & Syrgkanis, V. Orthogonal statistical learning. *The Annals of Statistics* **51**, 879–908 (2023).

[22] Künzel, S. R., Sekhon, J. S., Bickel, P. J. & Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* **116**, 4156–4165 (2019).

[23] Curth, A. & van der Schaar, M. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, 1810–1818 (2021).

[24] Kennedy, E. H. Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv:2004.14497* (2023).

[25] Nie, X. & Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108**, 299–319 (2021).

[26] Chernozhukov, V. *et al.* Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21**, C1–C68 (2018).

[27] Yoon, J., Jordon, J. & van der Schaar, M. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations* (2018).

[28] Evans, W. E. & Relling, M. V. Pharmacogenomics: Translating functional genomics into rational therapeutics. *Science* **286**, 487–491 (1999).

[29] Dang, L. E. *et al.* A causal roadmap for generating high-quality real-world evidence. *Journal of Clinical and Translational Science* **7**, E212 (2023).

[30] Petersen, M. L. & van der Laan, M. J. Causal models and learning from data: Integrating causal modeling and statistical estimation. *Epidemiology* **25**, 418–426 (2014).

[31] Hirano, K. & Imbens, G. W. The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, 73–84 (John Wiley & Sons, Chichester, UK, 2004).

[32] Specht, L. *et al.* Modern radiation therapy for hodgkin lymphoma: field and dose guidelines from the international lymphoma radiation oncology group (ilrog). *International Journal of Radiation Oncology, Biology, Physics* **89**, 854–862 (2014).

[33] van Geloven, N. *et al.* Prediction meets causal inference: The role of treatment in clinical prediction models. *European Journal of Epidemiology* **35**, 619–630 (2020).

[34] Imbens, G. W. & Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences* (Cambridge University Press, New York, NY, 2015).

[35] Chen, J., Vargas-Bustamante, A., Mortensen, K. & Ortega, A. N. Racial and ethnic disparities in health care access and utilization under the Affordable Care Act. *Medical Care* **54**, 140–146 (2016).

[36] Cinelli, C., Forney, A. & Pearl, J. *A crash course in good and bad controls*. Sociological Methods & Research (SAGE, Los Angeles, CA, 2022).

[37] Laffers, L. & Mellace, G. Identification of the average treatment effect when sutva is violated. *Working Paper* (2020).

[38] Huber, M. & Steinmayr, A. A framework for separating individual-level treatment effects from spillover effects. *Journal of Business & Economic Statistics* **39**, 422–436 (2021).

[39] Syrgkanis, V. *et al.* Machine learning estimation of heterogeneous treatment effects with instruments. In *Advances in Neural Information Processing Systems* (2019).

[40] Frauen, D. & Feuerriegel, S. Estimating individual treatment effects under unobserved confounding using binary instruments. In *International Conference on Learning Representations* (2023).

[41] Lim, B. Forecasting treatment responses over time using recurrent marginal structural networks. In *Advances in Neural Information Processing Systems* (2018).

[42] Liu, R., Yin, C. & Zhang, P. Estimating individual treatment effects with time-varying confounders. In *IEEE International Conference on Data Mining (ICDM)*, 382–391 (2020).

[43] Li, R. *et al.* G-Net: A deep learning approach to G-computation for counterfactual outcome prediction under dynamic treatment regimes. In *Machine Learning for Health* (2021).

[44] Bica, I., Alaa, A. M., Jordon, J. & van der Schaar, M. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations* (2020).

[45] Liu, R., Hunold, K. M., Caterino, J. M. & Zhang, P. Estimating treatment effects for time-to-treatment antibiotic stewardship in sepsis. *Nature Machine Intelligence* **5**, 421–431 (2023).

[46] Melnychuk, V., Frauen, D. & Feuerriegel, S. Causal transformer for estimating counterfactual outcomes. In *International Conference on Machine Learning*, 15293–15329 (2022).

[47] Schulam, P. & Saria, S. Reliable decision support using counterfactual models. In *Advances in Neural Information Processing Systems* (2017).

[48] Vanderschueren, T., Curth, A., Verbeke, W. & van der Schaar, M. Accounting for informative sampling when learning to forecast treatment outcomes over time. In *International Conference on Machine Learning* (2023).

[49] Seedat, N., Imrie, F., Bellot, A., Qian, Z. & van der Schaar, M. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. In *International Conference on Machine Learning*, 19497–19521 (2022).

[50] Hess, K., Melnychuk, V., Frauen, D. & Feuerriegel, S. Bayesian neural controlled differential equations for treatment effect estimation. In *International Conference on Learning Representations* (2024).

[51] Hatt, T., Berrevoets, J., Curth, A., Feuerriegel, S. & van der Schaar, M. Combining observational and randomized data for estimating heterogeneous treatment effects. *arXiv:2202.12891* (2022).

[52] Colnet, B. *et al.* Causal inference methods for combining randomized trials and observational studies: A review. *arXiv:2011.08047* (2020).

[53] Kallus, N., Puli, A. M. & Shalit, U. Removing hidden confounding by experimental grounding. *Advances in Neural Information Processing Systems* (2018).

[54] van der Laan, M. J., Polley, E. C. & Hubbard, A. E. Super learner. *Statistical Applications in Genetics and Molecular Biology* **6**, Article 25 (2007).

[55] Zheng, W. & van der Laan, M. J. Cross-validated targeted minimum-loss-based estimation. In van der Laan, M. J. & Rose, S. (eds.) *Targeted learning: causal inference for observational and experimental data*, 459–474 (Springer, New York, NY, 2011).

[56] Díaz, I. & van der Laan, M. J. Targeted data adaptive estimation of the causal dose–response curve. *Journal of Causal Inference* **1**, 171–192 (2013).

[57] Luedtke, A. R. & van der Laan, M. J. Super-learning of an optimal dynamic treatment rule. *The International Journal of Biostatistics* **12**, 305–332 (2016).

[58] Athey, S. & Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* **113**, 7353–7360 (2016).

[59] Wager, S. & Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* **113**, 1228–1242 (2018).

[60] Athey, S., Tibshirani, J. & Wager, S. Generalized random forests. *The Annals of Statistics* **47**, 1148–1178 (2019).

[61] Shalit, U., Johansson, F. D. & Sontag, D. Estimating individual treatment effect: Generalization bounds and algorithms. In *International Conference on Machine Learning*, 3076–3085 (2017).

[62] Shi, C., Blei, D. & Veitch, V. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems* (2019).

[63] Bach, P., Chernozhukov, V., Kurz, M. S. & Spindler, M. DoubleML: An object-oriented implementation of double machine learning in Python. *Journal of Machine Learning Research* **23**, 2469–2474 (2022).

[64] Kennedy, E. H., Ma, Z., McHugh, M. D. & Small, D. S. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 1229–1245 (2017).

[65] Nie, L., Ye, M., Liu, Q. & Nicolae, D. VCNet and functional targeted regularization for learning causal effects of continuous treatments. In *International Conference on Learning Representations* (2021).

[66] Bica, I., Jordon, J. & van der Schaar, M. Estimating the effects of continuous-valued interventions using generative adversarial networks. In *Advances in Neural Information Processing Systems* (2020).

[67] Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20**, 217–240 (2011).

[68] Schwab, P., Linhardt, L., Bauer, S., Buhmann, J. M. & Karlen, W. Learning counterfactual representations for estimating individual dose-response curves. In *AAAI Conference on Artificial Intelligence* (2020).

[69] Schweisthal, J., Frauen, D., Melnychuk, V. & Feuerriegel, S. Reliable off-policy learning for dosage combinations. In *Advances in Neural Information Processing Systems* (2023).

[70] Melnychuk, V., Frauen, D. & Feuerriegel, S. Normalizing flows for interventional density estimation. In *International Conference on Machine Learning*, 24361–24397 (2023).

[71] Banerji, C. R., Chakraborti, T., Harbron, C. & MacArthur, B. D. Clinical AI tools must convey predictive uncertainty for each individual patient. *Nature Medicine* (2023).

[72] Alaa, A. M. & van der Schaar, M. Bayesian inference of individualized treatment effects using multi-task Gaussian processes. In *Advances in Neural Information Processing Systems* (2017).

[73] Alaa, A., Ahmad, Z. & van der Laan, M. Conformal meta-learners for predictive inference of individual treatment effects. In *Advances in Neural Information Processing Systems* (2023).

[74] Morzywołek, P., Decruyenaere, J. & Vansteelandt, S. On a general class of orthogonal learners for the estimation of heterogeneous treatment effects. *arXiv:2303.12687* (2023).

[75] Curth, A., Svensson, D., Weatherall, J. & van der Schaar, M. Really doing great at estimating CATE? A critical look at ML benchmarking practices in treatment effect estimation. In *Advances in Neural Information Processing Systems* (2021).

[76] Boyer, C. B., Dahabreh, I. J. & Steingrimsson, J. A. Assessing model performance for counterfactual predictions. *arXiv:2308.13026* (2023).

[77] Keogh, R. H. & van Geloven, N. Prediction under hypothetical interventions: Evaluation of performance using longitudinal observational data. *arXiv:2304.10005* (2023).

[78] Curth, A. & van der Schaar, M. In search of insights, not magic bullets: Towards demystification of the model selection dilemma in heterogeneous treatment effect estimation. In *International Conference on Machine Learning* (2023).

[79] Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y. & Van Der Laan, M. J. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research* **21**, 31–54 (2012).

[80] Jesson, A., Mindermann, S., Shalit, U. & Gal, Y. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems* (2020).

[81] Rudolph, K. E. *et al.* When effects cannot be estimated: Redefining estimands to understand the effects of naloxone access laws. *Epidemiology* **33**, 689–698 (2022).

[82] Cornfield, J. *et al.* Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer institute* **22**, 173–203 (1959).

[83] Frauen, D., Melnychuk, V. & Feuerriegel, S. Sharp bounds for generalized causal sensitivity analysis. In *Advances in Neural Information Processing Systems* (2023).

[84] Kallus, N., Mao, X. & Zhou, A. Interval estimation of individual-level causal effects under unobserved confounding. In *International Conference on Artificial Intelligence and Statistics*, 2281–2290 (2019).

[85] Jin, Y., Ren, Z. & Candès, E. J. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences* **120**, e2214889120 (2023).

[86] Dorn, J. & Guo, K. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association* forthcoming (2022).

[87] Oprescu, M. *et al.* B-learner: Quasi-oracle bounds on heterogeneous causal effects under hidden confounding. *arXiv:2304.10577* (2023).

[88] Sharma, A., Syrgkanis, V., Zhang, C. & Kıcıman, E. DoWhy: Addressing challenges in expressing and validating causal assumptions. In *Workshop on the Neglected Assumptions in Causal Inference (NACI) at the International Conference on Machine Learning* (2021).

[89] Vokinger, K. N., Feuerriegel, S. & Kesselheim, A. S. Mitigating bias in machine learning for medicine. *Communications Medicine* **1**, 25 (2021).

[90] Hernán, M. A. & Robins, J. M. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology* **183**, 758–764 (2016).

[91] Xu, J. *et al.* Protocol for the development of a reporting guideline for causal and counterfactual prediction models in biomedicine. *BMJ Open* **12**, e059715 (2022).

[92] Fournier, J. C. *et al.* Antidepressant drug effects and depression severity: A patient-level meta-analysis. *JAMA* **303**, 47–53 (2010).

[93] Booth, C. M., Karim, S. & Mackillop, W. J. Real-world data: towards achieving the achievable in cancer care. *Nature Reviews Clinical Oncology* **16**, 312–325 (2019).

[94] Chien, I. *et al.* Multi-disciplinary fairness considerations in machine learning for clinical trials. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 906–924 (2022).

[95] Ross, E. L. *et al.* Estimated average treatment effect of psychiatric hospitalization in patients with suicidal behaviors: A precision treatment analysis. *JAMA Psychiatry* (2023).

[96] Cole, S. R. & Stuart, E. A. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology* **172**, 107–115 (2010).

[97] Hatt, T., Tschernutter, D. & Feuerriegel, S. Generalizing off-policy learning under sample selection bias. In *Uncertainty in Artificial Intelligence*, 769–779 (2022).

[98] Sherman, R. E. *et al.* Real-world evidence—what is it and what can it tell us. *New England Journal of Medicine* **375**, 2293–2297 (2016).

[99] Norgeot, B. *et al.* Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. *Nature Medicine* **26**, 1320–1324 (2020).

[100] Von Elm, E. *et al.* The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *The Lancet* **370**, 1453–1457 (2007).

This version of the article has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1038/s41591-024-02902-1. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use https://www.springernature.com/gp/open-research/policies/acceptedmanuscript-terms