nature medicine

Perspective

https://doi.org/10.1038/s41591-024-02902-1

Causal machine learning for predicting treatment outcomes

Received: 3 January 2024

Accepted: 4 March 2024

Published online: 19 April 2024

Check for updates

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] & Mihaela van der Schaar[7,8]

Causal machine learning (ML) offers flexible, data-driven methods for predicting treatment outcomes including efficacy and toxicity, thereby supporting the assessment and safety of drugs. A key benefit of causal ML is that it allows for estimating individualized treatment effects, so that clinical decision-making can be personalized to individual patient profiles. Causal ML can be used in combination with both clinical trial data and real-world data, such as clinical registries and electronic health records, but caution is needed to avoid biased or incorrect predictions. In this Perspective, we discuss the benefits of causal ML (relative to traditional statistical or ML approaches) and outline the key components and steps. Finally, we provide recommendations for the reliable use of causal ML and effective translation into the clinic.

Assessing the effectiveness of treatments is crucial to ensure patient safety and personalize patient care. Recent innovations in 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, changes in patient outcomes due to treatment[1]. Causal ML can be used 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 enables estimation of individualized treatment effects, as well as personalized predictions of potential patient outcomes (for example, survival, readmission, quality of life or toxicity) under different treatment scenarios. 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. Still, cautious use is important as causal inference rests on formal assumptions that cannot be tested.

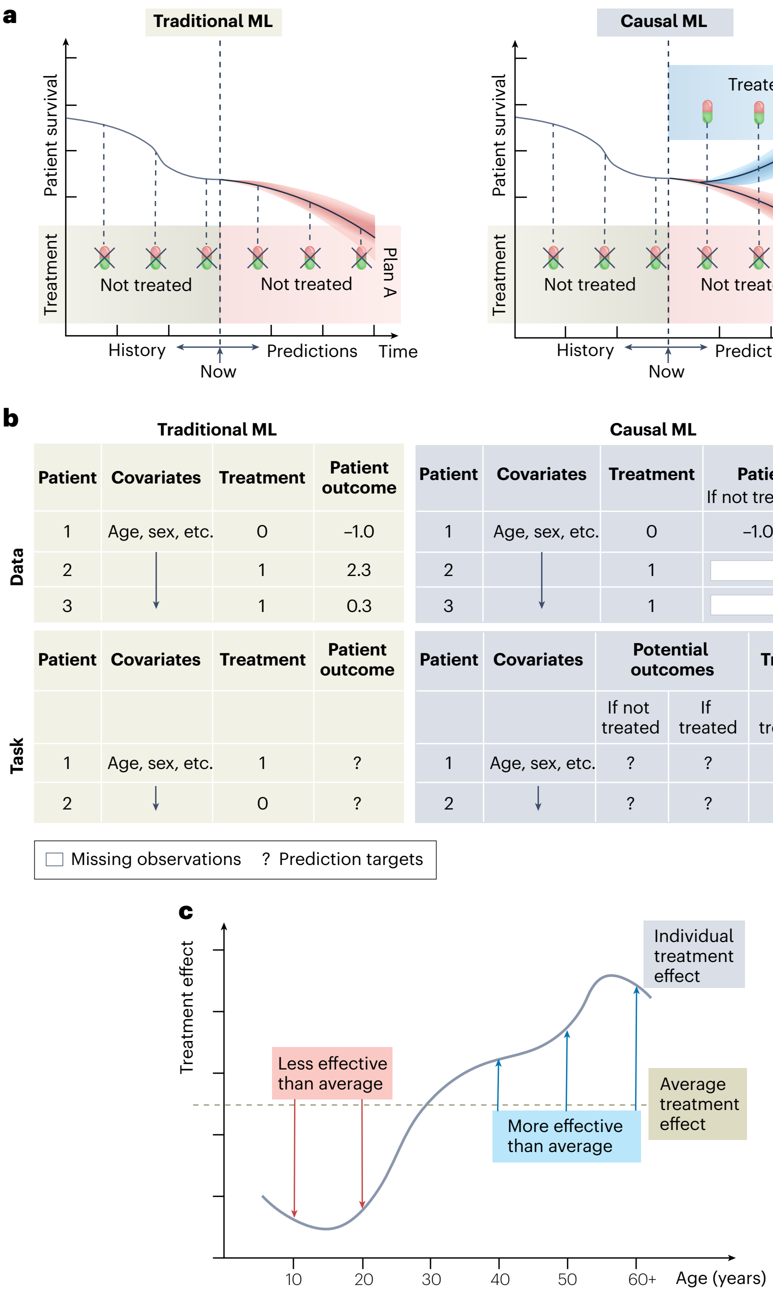
In this Perspective, we explain how causal ML differs from traditional statistical and ML approaches, and we discuss the essential components and steps for its use in the clinic. We provide recommendations for avoiding common technical pitfalls and outline a path to translation of this approach into clinical practice.

Causal ML in medicine

In medicine, causal ML offers several opportunities for estimating individualized 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 multimodal datasets containing images, text or time series, as well as genetic data. For example, one could estimate treatment effects from computed tomography 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[2]. For example, individual differences in drug metabolism can lead to serious side effects for drugs in some patients but can be lifesaving in others[3], so a causal ML approach 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 identify for which patient subgroups treatment is effective (Fig. 1c).

[1]LMU Munich, Munich, Germany. [2]Munich Center for Machine Learning, Munich, Germany. [3]Department of Applied Mathematics & Theoretical Physics, University of Cambridge, Cambridge, UK. [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, MA, USA. [7]Cambridge Centre for AI in Medicine, University of Cambridge, Cambridge, UK. [8]The Alan Turing Institute, London, UK. e-mail: feuerriegel@lmu.de

https://doi.org/10.1038/s41591-024-02902-1



Traditional ML

Data

Patient	Covariates	Treatment	Patient outcome
1	Age, sex, etc.	0	–1.0
2	↓	1	2.3
3		1	0.3

Task

Patient	Covariates	Treatment	Patient outcome
1	Age, sex, etc.	1	?
2	↓	0	?

Causal ML

Data

Patient	Covariates	Treatment	Patient outcome	
			If not treated	If treated
1	Age, sex, etc.	0	–1.0	□
2	↓	1	□	2.3
3		1	□	0.3

Task

Patient	Covariates	Potential outcomes		Treatment effect
		If not treated	If treated	If treated → If not treated
1	Age, sex, etc.	?	?	?
2	↓	?	?	?

□ Missing observations ? Prediction targets

Fig. 1 | Causal ML for predicting treatment outcomes. a, Different from traditional ML, causal ML aims to (i) estimate the treatment effect or (ii) predict the potential patient outcomes themselves owing to treatments. As such, one can perform 'what if' reasoning to evaluate how patient outcomes will change due to administering a treatment. **b**, Causal ML is challenging owing to the 'fundamental problem' of causal inference—in that not all potential outcomes can be observed and are thus missing in the data. Unless potential outcomes are explicitly needed, treatment effect estimation is preferred. **c**, Treatment effect heterogeneity refers to the variation in the response to treatment across different subgroups of a patient population (for example, according to age), indicating that the effectiveness of the intervention is not uniform for all individuals. For this, one must move beyond the ATE and obtain individualized treatment effects.

Despite these potential benefits, causal ML poses distinct challenges that necessitate custom methods. In addition, the appropriate application of this approach requires an understanding of how causal ML differs from traditional statistical and ML approaches.

When should I use causal 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[4], while causal ML quantifies changes in outcomes due to treatment, so that treatment effects can be estimated (Fig. 1a). A typical use case for traditional ML is risk scoring, such as predicting the probability of diabetes onset to understand which patients are at high risk—but without saying what the best treatment plan is[5–9]. By contrast, causal ML aims to answer 'what if' questions. For example, causal ML could estimate how the risk of diabetes onset will change if the patient receives an antidiabetic drug[10–12], so that decisions can be made about whether to administer such a 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 or longest duration of survival[13].

Methods for estimating treatment effects have a long tradition in the statistical literature (for example, refs. 14–17). Causal ML builds

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 ML that aims to estimate causal quantities (for example, ATE and CATE) or to predict 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 ref. 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 potential outcomes are clearly defined and observable in principle.

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 causal quantities such as treatment effects to 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.

SUTVA: The outcome for any patient does not depend on the treatment assignment of other patients, and there is no hidden variation in the effect of the treatment across different settings or populations.

Unconfoundedness: Given observed covariates, the treatment assignment is independent of the potential outcomes. This is the case, for example, when there are no unobserved confounders, that is, variables influencing both the treatment and the outcome. The assumption is also called ignorability.

BOX 2

Comparison of causal ML versus traditional statistics

Owing 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 ML (for example, refs. 23,27,49,53,54,61,98,99). 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 ML models[23,49,61].

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

upon the same problem setup but makes changes to the estimation strategy. Hence, the core benefit of 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 statistical literature (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. By 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.

The fundamental problem of causal inference

Estimating treatment effects from data requires custom methods. This is because treatment effects for individual patients are not observable owing to the so-called fundamental problem of causal inference[18,19]: that is, 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 (Fig. 1b). Therefore, the estimation of treatment effects or other causal quantities that are based on such unobserved outcomes poses challenges that do not exist in traditional, predictive ML.

First, to obtain a causal quantity (such as response to treatment) 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 treatment decisions and subsequent patient outcomes. If unmeasured confounding is present, the estimated treatment effects may suffer from confounding bias and, as a result, can be incorrect[20]. 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. This is because intervening on the treatment

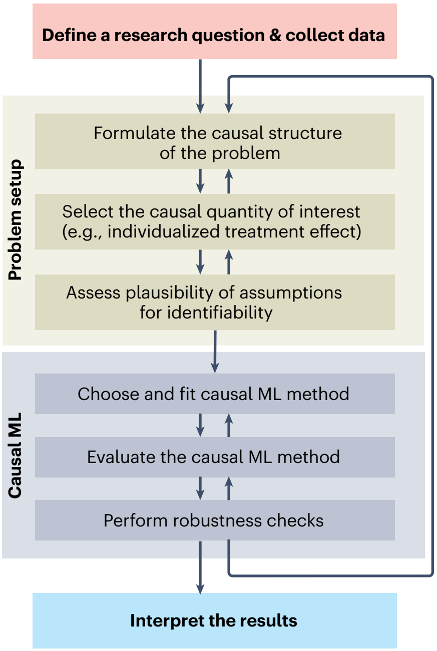

Define a research question & collect data
Problem setup
Formulate the causal structure of the problem
Select the causal quantity of interest (e.g., individualized treatment effect)
Assess plausibility of assumptions for identifiability
Causal ML
Choose and fit causal ML method
Evaluate the causal ML method
Perform robustness checks
Interpret the results

Fig. 2 | Workflow for causal ML in medicine. To predict treatment outcomes, assumptions on the causal structure of the problem must be made. This is relevant regardless of whether causal ML approaches or traditional statistical approaches are used. Subsequently, the causal quantity of interest can be predicted by causal ML.

variable could also affect other patient characteristics. As an example, consider a patient with a high body mass index whose doctor recommends quitting smoking, and for whom the diabetes risk should be predicted. Literature from traditional ML would suggest using both the body mass index and smoking behavior to predict the diabetes risk under a smoking versus no-smoking scenario; however, this approach would ignore that stopping smoking would also change a patient's body mass index. To address this issue, ML needs to be embedded in a causal framework.

The causal ML workflow

The process of predicting treatment outcomes with causal ML can be broken down into a few key steps (Fig. 2), which are discussed in the sections below. Following this workflow[21,22] should help researchers to clearly define the research question and then guide their formulation of the problem structure, their choice of the causal quantity of interest, the causal ML method, the evaluation metric and the appropriate robustness checks to validate the reliability of the estimates.

Formulate the causal structure of the problem

To estimate the effectiveness of treatments, information about the following variables is necessary[19]: the treatment of interest, the observed patient outcome and patient characteristics (covariates) such as age, gender and the medical history. 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[19], the variables can influence each other as shown by the 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, such as clinical registries and electronic health records, the treatment assignment follows some typically unknown procedure, depending on the patient characteristics. For example, patients with a very severe illness are likely to get a more aggressive form of treatment, implying that the patient characteristics differ across treatment groups. This contrasts with RCTs, where treatments are randomized and, as a result, the patient characteristics are similar across treatment groups. This is captured by the propensity score, which is the probability of receiving a treatment given the patient covariates[14]. In RCTs, the propensity score is known (for example, the propensity score is 50% in completely randomized trials with two treatment arms of equal size). By contrast, the propensity score in RWD is unknown, but it can be estimated to account for differences in the patient populations.

Select the causal quantity of interest

Causal quantities, such as the response to treatment, are commonly formalized based on the 'potential outcomes framework'[15]. The framework conceptualizes 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 (Fig. 3b); the degree of effect heterogeneity and 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 predicted by the causal ML method.

Degree of effect heterogeneity. Traditionally, the average treatment effect (ATE) is widely used in clinical trials. The ATE measures effects at the level of the study population[14]. By comparing the average patient outcome for those receiving the treatment versus those who do not (control group), the ATE helps in understanding how effective a treatment is, on average, across a specific patient cohort[23]. 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, the ATE cannot offer granular insights into whether patients with specific covariates may particularly benefit from a treatment, even though such heterogeneity in treatment effects can be of high interest in clinical practice (Fig. 1c). For a more granular view, 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 might even be harmful, which is relevant for individualizing treatment recommendations to specific patients.

Treatment type. Binary (discrete) treatments refer to a type of treatment variable that is dichotomous and thus has only two (or more) categories—for example, when answering questions of whether to treat or not to treat. By contrast, 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[24]. 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[25]. For continuous treatments, the treatment effectiveness is often also summarized by dose–response curves.

Individual patient outcomes. Besides the above, some applications in medicine are also interested in predicting the individual patient outcomes. Predicting patient outcomes is different from treatment effects, as the former gives granular predictions of the potential outcomes under different treatments, while the latter estimates only comparative changes in outcomes but not the outcomes themselves. Therefore, treatment effects primarily tell the advantages of one treatment over another, while potential outcomes can support decision-making

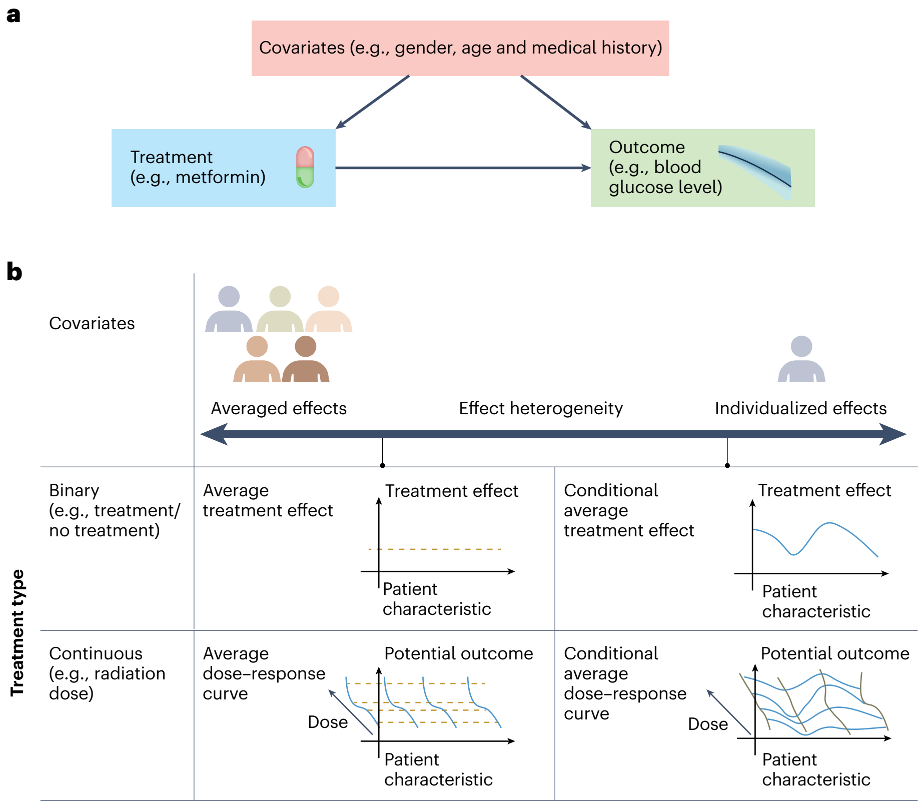


Fig. 3 | Formalizing tasks for causal ML. a, A causal graph must be assumed such as the example here. The arrows indicate the causal relationships between different variables. In the example, the assumption is made that patient outcomes are influenced by treatment and covariates. Note that the causal 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 estimand. The estimand can vary by the effect heterogeneity (average versus individualized) and treatment type (binary versus continuous). Depending on the causal graph and the causal quantity of interest, an appropriate causal ML method must be chosen.

in routine care by helping clinicians 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[26]. For example, while the treatment effect may say that a drug can improve the 5-year mortality by five percentage points, the predicted outcomes could inform us that the mortality is 15% with treatment and 20% without. However, in practice, the estimation of ATE and CATE is often an easier task than predicting potential outcomes[27] and, hence, is preferred when it is sufficient for decision-making.

Assess the plausibility of 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[19]. 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 a necessary step because, otherwise, it is impossible to estimate a treatment effect without bias, even with infinite data[19].

RCTs ensure the identifiability of treatment effects through fully randomized treatment assignment. However, treatment assignment in RWD is not fully randomized and depends on covariates, so that formal assumptions must be made[14]. The exact set of assumptions depends on which type of treatment effect is chosen. For the treatment effects discussed above, in addition to having independent and identically distributed data, three 'causal' assumptions are standard[14,28]. First, stable unit treatment value assumption (SUTVA) requires that the potential outcome coincides with the observed outcome for a given treatment and that the observed potential outcome on one patient should be unaffected by the particular assignment of treatments to other patients. This assumption implies that there is no interference whereby treating one patient influences the outcomes for another patient in the study population (for example, due to spillover or peer effects). The SUTVA assumption also implies that there is hidden variation in the treatment effect across hospitals or populations. SUTVA is also known as consistency assumption together with non-interference. Second, positivity (also called overlap) requires a nonzero probability of receiving a treatment. Positivity implies that, for each possible combination of patient characteristics, we can observe both treated and untreated patients. And third, 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—in other words, variables that influence both the treatment and the outcome. For example, unconfoundedness may be violated if patients with certain sociodemographic characteristics (such as race or income level) tend to have better access to treatments[29], and where the reason is not captured in the data. In principle, unconfoundedness can be addressed by capturing all relevant factors driving treatment assignment in RWD[30], yet it is generally challenging to validate this in practice. If confounders are not observed or not modeled (or even not known), then the estimated treatment effect might be biased and thus incorrect[20].

Importantly, assumptions such as those above are required for consistently estimating treatment effects from data, regardless of whether a causal ML approach or a traditional statistical approach is used. A natural challenge comes from the fact that assessing the plausibility of the assumptions is often difficult. Later, we discuss potential

BOX 3

Model-agnostic methods for CATE estimation

There are different ways in which meta-learners can leverage the data in a supervised learning setting for CATE estimation.

Plug-in learners: One approach 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[53]). Another way is to train two separate ML models for each treatment (called T-learner[53]). 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[27]. 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[27] and the so-called R-learner[98], which come with certain robustness guarantees[61,98,99].

The above meta-learners have different advantages and disadvantages. Unfortunately, there are no clear rules for choosing meta-learners but only high-level recommendations[54,75,100].

strategies to check the credibility of whether the assumptions hold. Notwithstanding, problem setups with alternative designs also exist. For example, some problem setups allow for relaxations of the SUTVA assumption (for example, by allowing for spillover effects)[31,32]. There also exist alternatives to assuming unconfoundedness in specific settings, such as through the use of instrumental variables[33,34]. Finally, there are problem setups that are not static but time-varying, so that a sequence of treatment decisions is made over time[35–44]. Researchers are also developing ways to effectively combine both observational and experimental data[45–47].

Choose and fit the causal ML method

There are different causal ML methods, which vary based on which causal graph and which causal quantity of interest is addressed. For example, a large body of literature focuses on causal ML for ATEs[23,48–52]. Here, a prominent method is based on so-called targeting to obtain an estimator that satisfies a semi-parametric efficient estimating equation. For CATE estimation with binary treatments, there are two broader categories of methods. On the one hand, so-called meta-learners[53] (Box 3) are model-agnostic methods for CATE estimation that can be used for treatment effect estimation in combination with an arbitrary ML model of choice (for example, a decision tree or a neural network[54]). A key advantage of model-agnostic methods is that the underlying ML model can be chosen to flexibly handle clinical data sources such as electronic health records. 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 that are particularly useful for clinical application are the causal tree[55] and the causal forest[56,57], 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[58,59]. A different set of methods is needed for predicting the response to continuous treatment variables—for settings in which the intensity, dosage or level of exposure to a treatment can be flexibly chosen[24,60–67]. This is because the number of treatment values is infinite and not every value is observed in the data—making treatment effect estimation particularly challenging in this context.

Existing causal ML methods often generate point estimates. This can be a serious limitation in medical applications[68], where uncertainty estimates such as standard errors or confidence intervals are crucial for reliable decision-making[69]. However, there is also some progress. For example, for CATE estimation, the causal forest[56,57] is a method that offers rigorous uncertainty estimates. In addition, several other strategies have been developed recently, such as Bayesian methods[70] and conformal prediction[71], but still more research is needed.

Evaluate the causal ML method

Arguably, the best way to evaluate causal ML methods is to assess the accuracy in predicting patient outcomes from randomized data. While this does not allow assessment of treatment effects for 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. By contrast, benchmarking for the purpose of model selection is challenging, as both counterfactuals and ground-truth values of treatment effects are unknown[72–74]. As a remedy, two strategies are common. A simple strategy 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 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—that is, the treatment effect—is not evaluated. Another approach is to use pseudo-outcomes[75]. 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, this approach depends on the performance of the secondary model for pseudo-outcomes and tends to favor certain methods[75]. Overall, both strategies are merely heuristics and there is no 'perfect' solution.

Perform robustness checks

To validate the robustness of the treatment effect estimates against explicit violations of the different assumptions, so-called refutation methods are used[76]. Common refutation methods include adding a random variable to check if the treatment effect estimates remain consistent (as such a 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 semisynthetic 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 causal ML methods highly depends on the specific problem setting and should be carefully chosen and implemented. Even when the refutation methods yield a positive result, this is no guarantee that the assumptions are satisfied. Nevertheless, robustness checks that are best practice in ML are still essential—for example, to mitigate the risk of bias[77]—especially as the results in treatment effect estimation may heavily depend on both the data and the model choice.

Technical recommendations

To ensure the careful and reliable use of causal ML in clinical practice, we make several technical recommendations.

Checking the plausibility of assumptions

Assessing the plausibility of the underlying assumptions is crucial for the validity of treatment effect estimates, yet it is 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 that they are not too small or too large; otherwise, there may not be enough support in the data for reliable inferences[78]. 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[79]. If the positivity assumption is violated, one strategy is to exclude certain subgroups from the analysis as no reliable inferences for them can be made[78,80].

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[30]. An alternative is to adopt an instrumental variable approach[33,34]; but appropriate 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[81]. 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. Restrictions on the amount of confounding are based on domain expertise, typically by making comparisons to known, important causes that act as baselines (for example, risk factors such as age). Recently, a series of causal ML methods have been proposed that provide sharp bounds[82–86]. 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[20].

Reporting

Findings should be interpreted and reported 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 (for example, owing 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, owing 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[87,88]. 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 substantial 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.

Clinical 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 compared to a placebo varies substantially and tends to increase with baseline severity of the depression[89]. However, RCTs typically compare patient outcomes across two (or more) treatment arms, which would return the ATE at the population level, and the use of causal ML may help to define inclusion criteria for clinical trials or to identify predictive biomarkers (for example, certain genetic mutations in a tumor).

Furthermore, causal ML may offer flexible, data-driven methods to analyze treatment effect heterogeneity in RWD, including clinical registries and electronic health records. This is relevant as RCTs can be subject to limitations[90]; for example, costs may be prohibitive or treatment randomization can be unethical for vulnerable populations (for example, pregnant women)[91]. RWD together with causal ML could allow the estimation of 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[92]. Likewise, patient populations in RCTs are often not representative of the broader population[93], but one can account for this through causal ML[94] to better understand the post-approval efficacy of treatments. However, while the potential of RWD has been widely recognized[90,95], 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 ATE. To ensure patient safety, regulatory bodies could also assess how the treatment effect varies across different subpopulations, which would involve the CATE. Likewise, the CATE may help to identify subpopulations that are particularly responsive to a treatment (for example, 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 and predicting individual 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[69]. For example, point estimates might indicate substantial effect heterogeneity, especially in settings with limited data, while in fact there may be 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 that account for practical needs in medicine (for example, 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, consensus-based, tailored checklists for reporting and quality will need to be developed. While there are checklists for traditional, predictive ML[96] and for generating real-world evidence[88,97], 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 causal ML methods must undergo regulatory review for approval.

So far, research in causal ML has primarily evaluated the performance of different methods through simulations (for example, refs. 35, 37,38,40,42,44). However, simulations involve (semi)synthetic datasets that do not fully capture the nuances of real-world disease dynamics. Hence, generating clinical insights 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 trial approaches. For this, settings where clear guidelines are missing could be appropriate, so that causal ML can provide input to augment the decision-making of clinical professionals. Causal ML for predicting treatment outcomes requires both methodological knowledge as well as domain knowledge of disease dynamics; therefore, cross-disciplinary collaboration between ML experts and clinicians is crucial for developing tools for clinical use. 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 the possibility to draw novel conclusions about the efficacy and safety of treatments and to personalize treatment strategies, thus improving patient health. However, in practice, several challenges arise, not least ensuring the reliability and robustness of these methods. Successful examples of causal ML in clinical use are still lacking, so proof-of-concept studies involving cautious use in clinical practice should be prioritized as an important first step.

References

1. Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J. & Silva, R. Causal machine learning: a survey and open problems. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2206.15475 (2022).
2. Yoon, J., Jordon, J. & van der Schaar, M. GANITE: estimation of individualized treatment effects using generative adversarial nets. In *Proc. 6th International Conference on Learning Representations* (ICLR, 2018).
3. Evans, W. E. & Relling, M. V. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* **286**, 487–491 (1999).
4. Esteva, A. et al. A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
5. Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A. & Stiglic, G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci. Rep.* **10**, 11981 (2020).
6. 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).
7. Cahn, A. et al. Prediction of progression from pre-diabetes to diabetes: development and validation of a machine learning model. *Diabetes/Metab. Res. Rev.* **36**, e3252 (2020).
8. Zueger, T. et al. Machine learning for predicting the risk of transition from prediabetes to diabetes. *Diabetes Technol. Ther.* **24**, 842–847 (2022).
9. Krittanawong, C. et al. Machine learning prediction in cardiovascular diseases: a metaanalysis. *Sci. Rep.* **10**, 16057 (2020).
10. 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. *Lancet Diabetes Endocrinol.* **11**, 644–656 (2023).
11. Deng, Y. et al. Comparative effectiveness of second line glucose lowering drug treatments using real world data: emulation of a target trial. *BMJ Med.* **2**, e000419 (2023).
12. Kalia, S. et al. Emulating a target trial using primary-care electronic health records: sodium glucose cotransporter 2 inhibitor medications and hemoglobin A1c. *Am. J. Epidemiol.* **192**, 782–789 (2023).
13. 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 Netw. Open* **3**, e200452 (2020).
14. Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688–701 (1974).
15. Rubin, D. B. Causal inference using potential outcomes: design, modeling, decisions. *J. Am. Stat. Assoc.* **100**, 322–331 (2005).
16. Robins, J. M. Correcting for non-compliance in randomized trials using structural nested mean models. *Commun. Stat.* **23**, 2379–2412 (1994).
17. Robins, J. M. Robust estimation in sequentially ignorable missing data and causal inference models. In *1999 Proceedings of the American Statistical Association on Bayesian Statistical Science* 6–10 (2000).
18. Holland, P. W. Statistics and causal inference. *J. Am. Stat. Assoc.* **81**, 945–960 (1986).
19. Pearl, J. *Causality: Models, Reasoning, and Inference* (Cambridge University Press, 2009).
20. Hemkens, L. G. et al. Interpretation of epidemiologic studies very often lacked adequate consideration of confounding. *J. Clin. Epidemiol.* **93**, 94–102 (2018).
21. Dang, L. E. et al. A causal roadmap for generating high-quality real-world evidence. *J. Clin. Transl. Sci.* **7**, e212 (2023).
22. 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).
23. van der Laan, M. J. & Rubin, D. Targeted maximum likelihood learning. *Int. J. Biostatistics* **2**, 11 (2006).
24. Hirano, K. & Imbens, G. W. in *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family* (eds Gelman, A. & Meng, X.-L.) Ch. 7 (John Wiley & Sons, 2004).
25. Specht, L. et al. Modern radiation therapy for Hodgkin lymphoma: field and dose guidelines from the international lymphoma radiation oncology group (ILROG). *Int. J. Radiat. Oncol. Biol. Phys.* **89**, 854–862 (2014).
26. van Geloven, N. et al. Prediction meets causal inference: the role of treatment in clinical prediction models. *Eur. J. Epidemiol.* **35**, 619–630 (2020).
27. Kennedy, E. H. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electron. J. Stat.* **17**, 3008–3049 (2023).
28. Imbens, G. W. & Rubin, D. B. *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press, 2015).

29. 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. *Med. Care* **54**, 140–146 (2016).
30. Cinelli, C., Forney, A. & Pearl, J. A crash course in good and bad controls. *Sociol. Methods Res.* https://doi.org/10.1177/00491241221099552 (2022).
31. Laffers, L. & Mellace, G. *Identification of the average treatment effect when SUTVA is violated. Department of Economics SDU. Discussion Papers on Business and Economics No. 3* (University of Southern Denmark, 2020).
32. Huber, M. & Steinmayr, A. A framework for separating individual-level treatment effects from spillover effects. *J. Bus. Econ. Stat.* **39**, 422–436 (2021).
33. Syrgkanis, V. et al. Machine learning estimation of heterogeneous treatment effects with instruments. In *Proc. 33rd International Conference on Neural Information Processing Systems* (eds Wallach, H. M. & Larochelle, H.) 15193–15202 (NeurIPS, 2019).
34. Frauen, D. & Feuerriegel, S. Estimating individual treatment effects under unobserved confounding using binary instruments. In *Proc. 11th International Conference on Learning Representations* (ICLR, 2023).
35. Lim, B. Forecasting treatment responses over time using recurrent marginal structural networks. In *Proc. Advances in Neural Information Processing Systems 31* (eds Bengio, H. et al.) (NeurIPS, 2018).
36. Liu, R., Yin, C. & Zhang, P. Estimating individual treatment effects with time-varying confounders. In *Proc. IEEE International Conference on Data Mining (ICDM)* 382–391 (IEEE, 2020).
37. Li, R. et al. G-Net: a deep learning approach to G-computation for counterfactual outcome prediction under dynamic treatment regimes. In *Proc. Machine Learning for Health* (eds Roy, S. et al.) 282–299 (PMLR, 2021).
38. Bica, I., Alaa, A. M., Jordon, J. & van der Schaar, M. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *Proc. 8th International Conference on Learning Representations* 11790–11817 (ICLR, 2020).
39. Liu, R., Hunold, K. M., Caterino, J. M. & Zhang, P. Estimating treatment effects for time-to-treatment antibiotic stewardship in sepsis. *Nat. Mach. Intell.* **5**, 421–431 (2023).
40. Melnychuk, V., Frauen, D. & Feuerriegel, S. Causal transformer for estimating counterfactual outcomes. In *Proc. 39th International Conference on Machine Learning* (eds Chaudhuri, K. et al.) 15293–15329 (PMLR, 2022).
41. Schulam, P. & Saria, S. Reliable decision support using counterfactual models. In *Proc. 31st International Conference on Neural Information Processing Systems* (eds von Luxburg, U. et al.) 1696–1706 (NeurIPS, 2017).
42. Vanderschueren, T., Curth, A., Verbeke, W. & van der Schaar, M. Accounting for informative sampling when learning to forecast treatment outcomes over time. In *Proc. 40th International Conference on Machine Learning* (eds Krause, A. et al.) 34855–34874 (PMLR, 2023).
43. Seedat, N., Imrie, F., Bellot, A., Qian, Z. & van der Schaar, M. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations. In *Proc. 39th International Conference on Machine Learning* (eds Chaudhuri, K. et al.) 19497–19521 (PMLR, 2022).
44. Hess, K., Melnychuk, V., Frauen, D. & Feuerriegel, S. Bayesian neural controlled differential equations for treatment effect estimation. In *Proc. 12th International Conference on Learning Representations* (ICLR, 2024).
45. Hatt, T., Berrevoets, J., Curth, A., Feuerriegel, S. & van der Schaar, M. Combining observational and randomized data for estimating heterogeneous treatment effects. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2202.12891 (2022).
46. Colnet, B. et al. Causal inference methods for combining randomized trials and observational studies: a review. *Stat. Sci.* **39**, 165–191 (2024).
47. Kallus, N., Puli, A. M. & Shalit, U. Removing hidden confounding by experimental grounding. In *Proc. 32nd Conference on Neural Information Processing Systems* (eds Bengio, S. et al.) 10888–10897 (NeurIPS, 2018).
48. van der Laan, M. J., Polley, E. C. & Hubbard, A. E. Super learner. *Stat. Appl. Genet. Mol. Biol.* **6**, 25 (2007).
49. van der Laan, M. J. & Rose, S. *Targeted Learning: Causal Inference for Observational and Experimental Data* 1st edn (Springer, 2011).
50. Zheng, W. & van der Laan, M. J. in *Targeted Learning: Causal Inference for Observational and Experimental Data* 1st edn, 459–474 (Springer, 2011).
51. Díaz, I. & van der Laan, M. J. Targeted data adaptive estimation of the causal dose–response curve. *J. Causal Inference* **1**, 171–192 (2013).
52. Luedtke, A. R. & van der Laan, M. J. Super-learning of an optimal dynamic treatment rule. *Int. J. Biostatistics* **12**, 305–332 (2016).
53. Künzel, S. R., Sekhon, J. S., Bickel, P. J. & Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl Acad. Sci. USA* **116**, 4156–4165 (2019).
54. Curth, A. & van der Schaar, M. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *Proc. 24th International Conference on Artificial Intelligence and Statistics* (eds Banerjee, A. & Fukumizu, K.) 1810–1818 (PMLR, 2021).
55. Athey, S. & Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proc. Natl Acad. Sci. USA* **113**, 7353–7360 (2016).
56. Wager, S. & Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* **113**, 1228–1242 (2018).
57. Athey, S., Tibshirani, J. & Wager, S. Generalized random forests. *Ann. Stat.* **47**, 1148–1178 (2019).
58. Shalit, U., Johansson, F. D. & Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *Proc. 34th International Conference on Machine Learning* (eds Precup, D. & Teh, Y. W.) 3076–3085 (PMLR, 2017).
59. Shi, C., Blei, D. & Veitch, V. Adapting neural networks for the estimation of treatment effects. In *Proc. 33rd International Conference on Neural Information Processing Systems* (eds Wallach, H. M. et al.) 2496–2506 (NeurIPS, 2019).
60. Bach, P., Chernozhukov, V., Kurz, M. S. & Spindler, M. DoubleML: an object-oriented implementation of double machine learning in Python. *J. Mach. Learn. Res.* **23**, 2469–2474 (2022).
61. Foster, D. J. & Syrgkanis, V. Orthogonal statistical learning. *Ann. Stat.* **51**, 879–908 (2023).
62. Kennedy, E. H., Ma, Z., McHugh, M. D. & Small, D. S. Nonparametric methods for doubly robust estimation of continuous treatment effects. *J. R. Stat. Soc. Series B Stat. Methodol.* **79**, 1229–1245 (2017).
63. Nie, L., Ye, M., Liu, Q. & Nicolae, D. VCNet and functional targeted regularization for learning causal effects of continuous treatments. In *Proc. 9th International Conference on Learning Representations* (ICLR, 2021).
64. Bica, I., Jordon, J. & van der Schaar, M. Estimating the effects of continuous-valued interventions using generative adversarial networks. In *Proc. 34th Annual Conference on Neural Information Processing Systems* (eds Larochelle, H. et al.) (NeurIPS, 2020).
65. Hill, J. L. Bayesian nonparametric modeling for causal inference. *J. Computational Graph. Stat.* **20**, 217–240 (2011).
66. Schwab, P., Linhardt, L., Bauer, S., Buhmann, J. M. & Karlen, W. Learning counterfactual representations for estimating individual dose-response curves. In *Proc. 34th AAAI Conference on Artificial Intelligence* 5612–5619 (AAAI, 2020).

67. Schweisthal, J., Frauen, D., Melnychuk, V. & Feuerriegel, S. Reliable off-policy learning for dosage combinations. In *Proc. 37th Annual Conference on Neural Information Processing Systems* (NeurIPS, 2023).
68. Melnychuk, V., Frauen, D. & Feuerriegel, S. Normalizing flows for interventional density estimation. In *Proc. 40th International Conference on Machine Learning* (eds Krause, A. et al.) 24361–24397 (PMLR, 2023).
69. Banerji, C. R., Chakraborti, T., Harbron, C. & MacArthur, B. D. Clinical AI tools must convey predictive uncertainty for each individual patient. *Nat. Med.* **29**, 2996–2998 (2023).
70. Alaa, A. M. & van der Schaar, M. Bayesian inference of individualized treatment effects using multi-task Gaussian processes. In *Proc. 31st Annual Conference on Neural Information Processing Systems* (eds von Luxburg, U. et al.) 3425–3433 (NeurIPS, 2017).
71. Alaa, A., Ahmad, Z. & van der Laan, M. Conformal meta-learners for predictive inference of individual treatment effects. In *Proc. 37th Annual Conference on Neural Information Processing Systems* (eds Oh, A. et al.) (NeurIPS, 2023).
72. 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 *Proc. 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (eds Vanschoren, J. & Yeung, S.-K.) (NeurIPS, 2021).
73. Boyer, C. B., Dahabreh, I. J. & Steingrimsson, J. A. Assessing model performance for counterfactual predictions. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2308.13026 (2023).
74. Keogh, R. H. & van Geloven, N. Prediction under interventions: evaluation of counterfactual performance using longitudinal observational data. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2304.10005 (2023).
75. 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 *Proc. 40th International Conference on Machine Learning* (eds Krause, A. et al.) 6623–6642 (PMLR, 2023).
76. Sharma, A., Syrgkanis, V., Zhang, C. & Kıcıman, E. DoWhy: addressing challenges in expressing and validating causal assumptions. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2108.13518 (2021).
77. Vokinger, K. N., Feuerriegel, S. & Kesselheim, A. S. Mitigating bias in machine learning for medicine. *Commun. Med.* **1**, 25 (2021).
78. Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y. & van der Laan, M. J. Diagnosing and responding to violations in the positivity assumption. *Stat. Methods Med. Res.* **21**, 31–54 (2012).
79. Jesson, A., Mindermann, S., Shalit, U. & Gal, Y. Identifying causal-effect inference failure with uncertainty-aware models. In *Proc. 34th Conference on Neural Information Processing Systems* (eds Larochelle, H. et al.) 11637–11649 (NeurIPS, 2020).
80. 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).
81. Cornfield, J. et al. Smoking and lung cancer: recent evidence and a discussion of some questions. *J. Natl Cancer Inst.* **22**, 173–203 (1959).
82. Frauen, D., Melnychuk, V. & Feuerriegel, S. Sharp bounds for generalized causal sensitivity analysis. In *Proc. 37th Annual Conference on Neural Information Processing Systems* (eds Oh, A. et al.) (NeurIPS, 2023).
83. Kallus, N., Mao, X. & Zhou, A. Interval estimation of individual-level causal effects under unobserved confounding. In *Proc. 22nd International Conference on Artificial Intelligence and Statistics* (eds Chaudhuri, K. & Sugiyama, M.) 2281–2290 (PMLR, 2019).
84. Jin, Y., Ren, Z. & Candès, E. J. Sensitivity analysis of individual treatment effects: a robust conformal inference approach. *Proc. Natl Acad. Sci. USA* **120**, e2214889120 (2023).
85. Dorn, J. & Guo, K. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *J. Am. Stat. Assoc.* **118**, 2645–2657 (2023).
86. Oprescu, M. et al. B-learner: quasi-oracle bounds on heterogeneous causal effects under hidden confounding. In *Proc. 40th International Conference on Machine Learning* (eds Krause, A. et al.) 26599–26618 (PMLR, 2023).
87. Hernán, M. A. & Robins, J. M. Using big data to emulate a target trial when a randomized trial is not available. *Am. J. Epidemiol.* **183**, 758–764 (2016).
88. 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).
89. Fournier, J. C. et al. Antidepressant drug effects and depression severity: a patient-level meta-analysis. *JAMA* **303**, 47–53 (2010).
90. Booth, C. M., Karim, S. & Mackillop, W. J. Real-world data: towards achieving the achievable in cancer care. *Nat. Rev. Clin. Oncol.* **16**, 312–325 (2019).
91. Chien, I. et al. Multi-disciplinary fairness considerations in machine learning for clinical trials. In *Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency (FACCT '22)* 906–924 (ACM, 2022).
92. Ross, E. L. et al. Estimated average treatment effect of psychiatric hospitalization in patients with suicidal behaviors: a precision treatment analysis. *JAMA Psychiatry* **81**, 135–143 (2023).
93. Cole, S. R. & Stuart, E. A. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *Am. J. Epidemiol.* **172**, 107–115 (2010).
94. Hatt, T., Tschernutter, D. & Feuerriegel, S. Generalizing off-policy learning under sample selection bias. In *Proc. 38th Conference on Uncertainty in Artificial Intelligence* (eds Cussens, J. & Zhang, K.) 769–779 (PMLR, 2022).
95. Sherman, R. E. et al. Real-world evidence—what is it and what can it tell us. *N. Engl. J. Med.* **375**, 2293–2297 (2016).
96. Norgeot, B. et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat. Med.* **26**, 1320–1324 (2020).
97. Von Elm, E. et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Lancet* **370**, 1453–1457 (2007).
98. Nie, X. & Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108**, 299–319 (2021).
99. Chernozhukov, V. et al. Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21**, C1–C68 (2018).
100. Morzywołek, P., Decruyenaere, J. & Vansteelandt, S. On a general class of orthogonal learners for the estimation of heterogeneous treatment effects. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.12687 (2023).

Acknowledgements

S.F. acknowledges funding via Swiss National Science Foundation Grant 186932.

Author contributions

All authors contributed to conceptualization, manuscript writing and approval of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Stefan Feuerriegel.

Peer review information *Nature Medicine* thanks Matthew Sperrin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Karen O'Leary, in collaboration with the *Nature Medicine* team.

Reprints and permissions information is available at www.nature.com/reprints.