Using Machine Learning to Individualize Treatment Effect Estimation: Challenges and Opportunities

Alicia Curth[1,*], Richard W. Peck[2,3], Eoin McKinney[4,5], James Weatherall[6] and Mihaela van der Schaar[1,5,7]

The use of data from randomized clinical trials to justify treatment decisions for real-world patients is the current state of the art. It relies on the assumption that average treatment effects from the trial can be extrapolated to patients with personal and/or disease characteristics different from those treated in the trial. Yet, because of heterogeneity of treatment effects between patients and between the trial population and real-world patients, this assumption may not be correct for many patients. Using machine learning to estimate the expected conditional average treatment effect (CATE) in individual patients from observational data offers the potential for more accurate estimation of the expected treatment effects in each patient based on their observed characteristics. In this review, we discuss some of the challenges and opportunities for machine learning to estimate CATE, including ensuring identification assumptions are met, managing covariate shift, and learning without access to the true label of interest. We also discuss the potential applications as well as future work and collaborations needed to further improve identification and utilization of CATE estimates to increase patient benefit.

Healthcare professionals try to make treatment decisions about individual patients using the best available evidence. In such evidence-based medicine, the data informing these individual patient decisions often comes from randomized clinical trials (RCTs) in populations of patients who have the same disease as the patient needing treatment now. The underlying assumption is that the patient needing treatment is similar to those studied in the clinical trials and will respond in a similar manner. Yet, this assumption is often not correct and there would be great benefit in having better methods to estimate expected effect in each individual patient. Such methods could be developed either during trials of novel medicines or from real-world evidence generated after approval, and – assuming they provide a net benefit – accompany those medicines as they are approved and embedded in healthcare systems around the world.

The problems with clinical trial evidence

RCTs remain the founding cornerstone of evidence-based medicine. They routinely form the basis of clinical decision making and guide health policy development. Yet, this clinical evidence base has several fundamental limitations. RCTs constrain recruitment to a limited group of clinically homogenous patients, limiting the impact of confounding conditions but also our ability to extrapolate results to excluded patient groups. Older patients and those with comorbidities (populations that substantially overlap) are routinely excluded from RCTs,[1,2] yet, they are increasingly the very population we need to treat. Staggeringly, more than half of RCTs exclude at least 75% of the potentially treatable population.[3,4] Despite these limitations, the evidence generated from RCTs is routinely extrapolated to excluded patients as the only alternative is not to use the treatment in them at all.[5] Sometimes, this extrapolation can be supported by approaches such as bridging studies, real-world studies, or Bayesian borrowing, although on other occasions there is no additional evidence on which to base decision making.[6]

RCTs also allocate all patients to either the intervention or the control group (receiving standard of care, a comparator drug, or placebo treatment) and seek to identify a difference in the "average" treatment effect at this group level.[7] Thus, another major limitation is that they can only summarize average treatment effects across that group: they cannot estimate the *individual* treatment effect for any given patient. Heterogeneity of treatment effects (HTEs) are commonly present and commonly ignored in RCTs[8–11] and are often associated with uncorrected traits, such as sex, age, and ethnicity. A failure to understand how treatment effects vary in these subgroups is both a driver of health inequality and a missed opportunity to individualize therapy.

Finally, RCTs identify a "static" treatment effect in a given population at a given time. Both health care and the population being treated are, by contrast, constantly evolving and changing. Patient demographics change, comorbidity patterns change, and so do treatment effects.[12] However, once undertaken, it is too expensive

[1]Department of Applied Mathematics & Theoretical Physics, University of Cambridge, Cambridge, UK; [2]Department of Pharmacology & Therapeutics, University of Liverpool, Liverpool, UK; [3]Roche Pharma Research & Early Development (pRED), Roche Innovation Center, Basel, Switzerland; [4]Cambridge Institute for Immunotherapy & Infectious Disease, Jeffrey Cheah Biomedical Center, Cambridge Biomedical Campus, Addenbrooke's Hospital, Cambridge, UK; [5]Cambridge Centre for AI in Medicine, Cambridge, UK; [6]AstraZeneca R&D Data Science and Artificial Intelligence, Cambridge, UK; [7]The Alan Turing Institute, London, UK. *Correspondence: Alicia Curth (amc253@cam.ac.uk)

Received August 30, 2023; accepted December 7, 2023. doi:10.1002/cpt.3159

and time-consuming for RCTs to be repeated to update clinical evidence. The same is true for estimating when, as opposed to if, a treatment should be given or indeed to determine when a treatment should be stopped.

Although they have defined modern medicine, RCTs are demonstrably not representative of the real-world population that is being treated and cannot estimate individual treatment effects or re-evaluate them easily over time. However, there is an urgent need to develop additional methods that can generate clinical evidence where RCTs cannot or will not apply. We argue that novel machine learning (ML) methods can be used to overcome these major limitations, creating a real-world evidence base that can augment and complement the power of RCTs.

Precision medicine vs. Individualized medicine

Precision medicine approaches – defined as getting the right treatment to the right patient at the right time – are long established with approximately one third of new drugs being classed as personalized medicines due to their targeting of patient subgroups.[13] Even apparently individualized biomarker-based approaches, such as testing for gene mutations, are actually about restricting treatment to the subpopulation with the mutation who have the potential to respond and avoiding treatment of those with no potential to respond. Because this is a subgroup of individuals, this remains precision medicine but not yet individualized therapy and there is usually still substantial HTE within the biomarker positive subpopulation.[14] In some situations, it is possible to make more accurate individual-level predictions to guide therapy. A few drugs are suitable for therapeutic drug monitoring using population pharmacokinetic/pharmacodynamic (PK/PD) models to guide dose adjustment to achieve a target concentration associated with an increased chance of benefit and/or reduced chance of adverse effects. Physiologically-based PK modeling (PBPK) can also be used to estimate individual patient PK parameters and guide individualized dose selection to achieve target drug concentrations.[15,16] Quantitative systems pharmacology (QSP) models also bring the potential to guide individual treatment decisions delivering improved outcomes compared with current standards of care.[17] However, modeling methods, such as PK/PD and PBPK, cannot help with the decision of which treatment to use. Furthermore, for many diseases, treatment choices follow treatment protocols and guidelines derived from populations of patients, with no method of accurately estimating the expected effect in each individual patient to help individualize treatment choices. ML, with its ability to identify complex patterns in data, has the potential to change this and allow much wider use of data-driven individualized treatment selection.

Challenges and opportunities for ML in individualizing treatment effect estimation

Providing treatment decision support to healthcare professionals requires methods that can make reliable individual level forecasts of the expected effect of each treatment option, ideally at different doses, including that of no treatment. Yet, today's ML methods are not usually designed to handle the need to make forecasts of treatment effectiveness, in order to decide who to treat, when, and with what intervention. One of the main reasons for this is that ML methods are particularly good at learning about the *status quo* from existing data to ultimately make outcome predictions that are exactly in line with the current distribution of data characteristics, but usually not designed for tasks that involve the need to reason about interventions on the data generating distributions leading to counterfactual scenarios, like "What would be the outcome if I were to change the treatment policy currently in place?," which is exactly what is needed to infer the effects of treatments. Additionally, there is another nuanced difference in the intended outputs of the two; the goal of standard supervised ML problems is the prediction of individual outcomes, whereas the goal of individualized treatment effect estimation is the estimation of *differences* of outcomes across different treatment levels. With this in mind, we therefore discuss the challenges and opportunities for the use of ML for forecasting individual treatment responses in this paper. (Throughout, we use the word "forecasting" to signify that we are interested in a patient's future under different hypothetical interventions, whereas "prediction" relates to the future without considering intervention.) We do so with two goals: on the one hand, we highlight that the ML literature on individualized treatment effect estimation has grown rapidly over the last years[18] and would like to bring solutions already developed therein closer to ultimate end-users from the medical domain. On the other hand, we re-emphasize that this problem setting inherently provides additional challenges that many ML methods are not natively equipped to handle – making it an interesting problem for ML researchers to explore further from a technical perspective.

SETUP: ESTIMATING INDIVIDUALIZED TREATMENT EFFECTS FROM STATIC OBSERVATIONAL DATA

We consider a standard static treatment effect estimation problem setup from observational data as usually formalized within the potential outcomes framework.[19] We consider datasets of patients characterized by pretreatment covariates collected in X – for example, age, weight, gender, and other baseline measurements on record. In this context, we then wish to estimate the effect a treatment has on an outcome Y for patients with different characteristics that could be discrete (such as mortality) or continuous (for example, change in a clinical severity score). We assume for now the simplest setting where outcomes Y and patient characteristics X are measured only once. That is, there is no additional temporal structure to our data, and patients can either receive ($A = 1$) or not receive ($A = 0$) a treatment of interest; we discuss more complex setups later. Conceptually, when deciding whether a patient should receive this treatment, we would then like to compare a patient's potential outcome they would realize without treatment $Y(0)$ to their potential outcome they would realize if given treatment $Y(1)$; the difference $Y(1)-Y(0)$ is then their individual treatment effect.

Our main quantity of interest we consider is therefore the conditional average treatment effect (CATE), which is the expected difference between the two potential outcomes for an individual characterized by covariates $X = x$.

$$\tau(x) = E[Y(1) - Y(0)|X = x]$$

This precisely gives us the expected (additive) effect a treatment has on the outcome of interest for an individual. This effect is sometimes also referred to as an HTE or individualized treatment effect. The term *average* in CATE just refers to the fact that we consider the expected (average) effect for someone characterized by the precise covariate profile of interest; it is generally assumed that the covariates X are granular enough to characterize individuals rather than just subgroups, although this will depend on the range of covariates considered. For example, considering only a handful of covariates may allow estimation of treatment effects by population subgroup (by sex or ethnicity, for example). With a larger range of covariates, individual profiles may be identified, allowing "granular" differentiation of responses at the individual level.

Instead of assessing only the change in outcome between the two treatment conditions, we may also be interested in making individualized forecasts of (absolute) outcomes under the two treatment conditions:

$$\mu_a(x) = E[Y(a) | X = x]$$

That is, personalized forecasts of outcome with ($\mu_1(x)$) and without ($\mu_0(x)$) treatment. These may be of interest in addition to $\tau(x)$, for example, because an invasive treatment with a large effect $\tau(x)$ may not need to be given if baseline outcome expectation $\mu_0(x)$ is positive enough. For example, deciding to perform surgery to remove a kidney stone may be unnecessary if there is a high chance it will be spontaneously passed anyway. Note that due to the linearity of expectation, we have that $\tau(x) = \mu_1(x) - \mu_0(x)$; the problems of personalized forecasting of outcomes under different treatments and personalized treatment effect estimation are thus inherently linked.

Unfortunately, due to the fundamental problem of causal inference,[20] we can only ever observe one of the two potential outcomes: namely $Y(A)$, the one associated with the treatment $A \in \{0, 1\}$ actually received by a patient: if they were not treated, we have $A = 0$ so we observe $Y(0)$, or if they were treated, we have $A = 1$ and observe $Y(1)$. Information on counterfactuals, what would have happened to a patient had they received a different treatment, is thus generally unavailable in reality. Instead of the ideal tuple $(X, Y(0), Y(1))$, we thus observe datasets consisting of tuples $(X, A, Y = Y(A))$ in practice. Importantly, in this paper, we consider treatment effect estimation from observational data where treatments A are *not* assigned completely randomly (note that data from randomized experiments can be considered a special case of observational data, where the treatment assignment mechanism is known and randomized). As we discuss in the following section, we thus allow *some* situations where treatment assignment was influenced by patient characteristics (as long as these are recorded).

In the next four sections (see also Table 1), we discuss the resultant challenges and opportunities inherent to estimating $\mu_a(x)$ and $\tau(x)$. We first discuss identification of effects, then learning under distribution shifts, and learning without the label of interest. Finally, we discuss the need to move beyond the simple standard setup outlined in this section to be able to tackle further complexity encountered in practice.

PROBLEM 1: IDENTIFICATION (BEFORE ANY LEARNING STARTS!)

The challenge

Possibly the most crucial challenge inherent in learning to forecast in the presence of treatments appears before any actual learning starts, namely ensuring identifiability of interventional quantities from the observed data. That means answering the question whether the data at hand can be used at all to learn how to make forecasts under intervention, or whether it is too biased to do so. Essentially, there needs to be sufficient information in the observed data to tell us something also about the part of the data that is unobserved,[21] i.e. the counterfactuals. In our context, one thus needs to make strong and untestable assumptions on the relationship between the true underlying potential outcomes $Y(0)$ and $Y(1)$ and treatment assignment A; namely, that once we are controlling for what is measured in X, A does not depend on unobserved information that also influences $Y(a)$ and there are no hidden/unobserved confounders. This ensures that treatment assignment is at random when conditioning on observables X, so that there is ignorability with respect to the treatment assignment mechanism.[22] In randomized experiments with successful randomization, this assumption should hold by design.

Further, one generally needs to make assumptions guaranteeing sufficient randomness in assignment,[22,23] conceptually ensuring that we could observe each potential outcome for each possible patient characteristic X. That is, we need to guarantee that each patient characteristic X has non-zero probability of

Table 1 The main problems, challenges, and opportunities for the use of ML to estimate individual-level treatment effects

Problem	Challenge	Opportunities for ML
Identification	Ensuring identification assumptions are met	1. Enabling inclusion of higher dim. controls 2. Enabling learning from proxy variables 3. Relaxing assumptions on model specification
Forecasting outcomes for different treatments	Handling distribution shifts across treatment groups	1. Use of domain adaptation techniques 2. Use of methods more robust to shifts
Estimating the effect of treatments	Learning without the label of interest	1. Going beyond "virtual twin" approaches 2. Creating ML methods that are better targeted at personalized effects
Looking beyond the basic static setting	Dealing with all the complexities of real data	1. Building methods that are able to handle multiple sources of missingness 2. Provision of new benchmark tasks

ML, machine learning.

receiving either treatment (or, more formally, for treatments assigned with propensity $\pi(x) = P(A = 1|X = x)$, we need that $\epsilon < \pi(x) < 1 - \epsilon$ for some small value $\epsilon > 0$). This overlap, or positivity, assumption ensures that we could, in principle, nonparametrically learn about outcomes for each individual because there is a chance that we would observe either potential outcome for all types of patients.

This need to make untestable assumptions does not seem to appear in standard prediction (or supervised learning) problems usually considered in ML. This is because, in standard prediction problems, the objective is usually to predict under the same conditions that data were observed,[24,25] which is very different from forecasting under intervention on data-generating mechanisms, as is the setting for treatment effect estimation.

Opportunities for ML

On the one hand, we note that identification really is a data problem, and not a learning problem, so ensuring that all relevant information is measured – for example, by consulting with domain experts[26] – is a task to be taken care of before learning starts, and thus important irrespective of the learning method used. On the other hand, ML can indirectly help by (i) making these assumptions more likely to hold and (ii) relaxing the needed assumptions slightly. A popular strategy for trying to ensure ignorability is to take a "kitchen sink" approach and to include as much pretreatment information into X as possible, making the learning problem more challenging through the high dimensionality of X. Domain expertise is also important in this context to ensure that no variables are included that increase or induce bias[27,28] because this requires expert knowledge of the likely underlying causal structure. ML has had great successes in learning good representations from high-dimensional and unstructured data, thus its ability to learn flexible functions of high-dimensional and/or unconventional control variables, for example, clinical notes,[29] could indirectly help with this challenge. Additionally, ML has shown promise to be used with relaxed assumptions that allow reconstruction of confounders from proxy variables.[30] Note, however, that the positivity assumptions are progressively harder to satisfy once data become more high-dimensional and/or multi-modal.[31] ML can help here by learning lower-dimensional representations of the high-dimensional data in which positivity holds.[32–34] The conceptually simplest way of achieving this would be to use a supervised feature selection algorithm, which reduces the number of features used by the model by learning whether the observed outcomes depend on this feature. More sophisticated approaches would instead learn lower-dimensional representations or embeddings of the higher-dimensional original feature space. Note that, because such lower-dimensional representations are learned in a supervised manner (i.e., by learning whether the observed outcomes depend on this feature), individual outcomes would not be expected to vary along dropped dimensions – the correct level of granularity needed for individualization will thus be *learned by the model*. Finally, note that ML implicitly helps relaxing other assumptions. Treatment effect estimation through covariate adjustment usually relies on correct specification of the outcome model, i.e. knowing how covariates and treatment impact outcome, to ensure identification.[35] Being able to rely on flexible data-adaptive ML methods ensures that a suitable specification can be learned from the data instead of being prespecified.

PROBLEM 2: LEARNING UNDER DISTRIBUTION SHIFT

The challenge

If we can overcome the first challenge by taking all steps to ensure that identification is likely to hold, we are able to use observed data to forecast the potential outcomes under different treatments, essentially by learning to predict the observed outcomes, but not without further challenges. Recall that learning to predict from data conceptually works by empirical minimization of a loss or empirical maximization of a likelihood, which ultimately involves trading off model fit across observed instances. For example, one usually fits predictive models for continuous outcomes, be it a linear regression model or a neural network, by minimizing the mean squared error of the predictor across the observed (training) sample. However, when we fit a model $f_a(x)$ for the outcomes Y using the data from the treatment group with $A = a$, where a could indicate either treatment ($a = 1$) or control group ($a = 0$), to estimate the potential outcome $\mu_a(x)$ in this way *using observational data*, we do so on a population with characteristics distributed proportional to the observed treatment propensity $\pi(x)$. If treatments were not assigned completely randomly, this observational distribution of patient characteristics in each treatment group is not equal to the marginal (population) distribution of patient characteristics. Yet ultimately it is the marginal, not the observational, distribution of patient characteristics that is the target population of interest when we make forecasts of treatment outcomes at test or deployment time. This problem, an instance of covariate shift studied in the ML literature,[36] means that individuals with high propensity for treatment a are over-represented in the sample when fitting $f_a(x)$, so that they might get too much weight in the tradeoff of model fit across the population. This is known to be especially problematic when models are misspecified.[36–38]

Opportunities for ML

The problem of covariate shift has been studied at length in the ML literature on domain adaptation where all kinds of distribution shifts arise naturally.[39] Proposed solutions in this context range from more classical statistical importance weighted learning[36] to adversarial learning of representations that aim to reduce distributional discrepancies.[40] Many of these ideas have been incorporated into the recent ML literature on HTE estimation proposing sophisticated deep learning architectures that include novel elements to overcome such distribution shifts induced by non-uniform treatment assignment policies.[41–46] Conversely, note that because this covariate shift is particularly problematic when models are misspecified, modern ML may help with this challenge simply by allowing the use of *more flexible* models that are less likely to be as misspecified and hence do *not require to trade off* errors across the input space, as backed up by recent observations that modern ML methods are more robust to covariate shifts.[47–49]

PROBLEM 3: LEARNING WITHOUT ACCESS TO THE TRUE "LABEL" OF INTEREST

The challenge

Often, we may care about estimating the potential outcomes with functions $f_a(x)$ only as an intermediate product: they ultimately just help us in estimating the CATE as $f_1(x) - f_0(x)$. This approach underlies, for example, the popular virtual twins method[50] for discovering subgroups from clinical trial data. In an ideal world, however, the actual label of interest for the learning problem would be the unobserved $Y(1) - Y(0)$. That is, in standard prediction problems, machine learning is *usually* deployed to learn to predict label (i.e., outcome) Y from covariates X – but in the CATE estimation problem, the true label of interest is the potential outcomes difference $Y(1) - Y(0)$, not the individual outcomes separately, and we would like to use ML to learn the relationship between $Y(1) - Y(0)$ and X directly. Whether we target the potential outcomes individually to first estimate their expectations and then take the difference, the dominant approach in early papers on the topic,[50,51] or whether we were to target $Y(1) - Y(0)$ *directly*, can matter especially when the potential outcomes are a complex function of covariates X *while the treatment effect is not*. That is, if there is a lot of prognostic information, such as risk factors that influence outcomes identically regardless of treatment status[52] and relatively little predictive information (effect modifiers), this can mean that the hypothetical problem in which $Y(1) - Y(0)$ is available for learning would be much easier to solve than the two potential outcome regressions separately because much of the complexity cancels out once we consider $Y(1) - Y(0)$. Ideally, we might sometimes thus actually like to estimate CATE directly instead of the two potential outcomes separately. Relatedly, note that when estimating the potential outcomes (POs) separately, estimation errors can either accumulate or cancel out across the two PO predictions – so that, in finite samples, the model with the best fit in terms of PO is not necessarily the model with the best fit on the CATE. To see this, note that when comparing an estimator that makes small yet different errors in estimating each of the two POs to an estimator that makes the same yet larger error when estimating both POs, the latter can perform much better in estimating the CATE if its errors cancel out perfectly, whereas for the former the errors compound.

Opportunities for ML

Much of the recent ML literature on CATE estimation has shown that approaches that simply estimate the two potential outcomes separately and take CATE equal to their difference, i.e. approaches referred to as virtual twins[50] or S- and T-learner[53], may not be optimally *targeted* at CATE. Targeting of estimators toward the estimand of interest, however, has been shown to be crucial in the context of *(population) average* treatment effect estimation in the extensive literature on targeted learning.[54] One strand of the CATE estimation literature[53,55–58] has therefore shown the empirical and theoretical advantage of using multistage estimators that use pseudo-outcomes constructed using some nuisance parameter estimates that are used as *surrogates* for the unobserved potential outcomes contrast $Y(1) - Y(0)$, making use of the fact that learning CATE this way converges much faster than through estimation of the potential outcomes separately. Another strand of this literature has shown that a related effect can be achieved when retargeting the inherent inductive, or simplicity, biases of ML methods at the CATE, for example by parameterizing or regularizing the CATE explicitly in a model that also outputs potential outcome predictions separately.[59–61] Relative to approaches that output CATE estimates only, this approach would be more suited in applications where the absolute value of the potential outcomes is relevant for decision making in addition to the estimated treatment effect, as in the previously highlighted example with an invasive treatment that may not need to be given if the baseline outcome expectation is positive enough.

A similar tradeoff applies also to the problem of selecting between multiple available candidate CATE estimators. Relying only on the performance of different estimators in terms of their ability to predict factual outcomes may not be the optimal way of choosing between trained models, and more targeted approaches have been considered recently.[62–66]

SCOPING FUTURE RESEARCH AGENDAS: LOOKING BEYOND THE STANDARD SETTING

The challenge

Most of the ML literature has only considered a small slice of the problems involving treatment effect estimation in practice. In reality, the observed data are often even coarser relative to the true forecasting targets of interest. There could be many more than two treatment levels, such as multiple treatment options and/or doses, and treatment assignment may not be the only source of missingness; censoring, competing events, and informative sampling can also lead to additional missing outcome information. Thus, one needs to consider identification problems, distribution shifts and targeting of estimators in much more generality for a much broader class of problems.

Opportunities for ML

Solutions for some specific problems with extended characteristics have been considered in the ML literature recently, for example, in the context of HTE estimation with missingness in covariates,[67] more complex treatments,[68–70] data from multiple sources,[71] (static) survival data with censoring,[72,73] or competing risks,[74] and regular (discrete-time) longitudinal data[75–77] or irregular[78,79] and even informatively sampled,[80] continuous-time, time-series data. However, unified treatment of settings with more general or unified missingness patterns is still an open problem in this literature, leaving ample room for future method development taking into account the sometimes very sparse nature of such data when multiple sources of missingness are present simultaneously in practice.

We believe that one of the reasons for lack of more general consideration of broader treatment effect estimation problems is that the ML literature often inherently relies on the availability of good benchmark datasets.[81] Test datasets are used to check whether proposed methods perform well in practical scenarios and thus provide proof-of-concept evidence for their usefulness. In the context of treatment effect estimation, the main benchmark datasets that are available consider the static binary setup, some of which are themselves inherently problematic.[82] Thus, good testbeds for any

other problems are still lacking and may encourage future method development. There is therefore an opportunity for the clinical pharmacology and other applied communities to define the research agenda in this context by providing guidance to the machine learning community on important problems to consider and ways of capturing success therein, such as by providing datasets and associated metrics that could be used for future method development.

Finally, there has been some discussion in the literature recently on the topic of combining ML and modeling, constituting another important avenue for future research. For example, the January 2022 issue of the *Journal of Pharmacokinetics and Pharmacodynamics* included several articles on the general theme of using ML to assist in developing QSP models. There are also examples of integrating a model into a ML framework to enable better forecasts of treatment effects[83] or guidance for dose adjustments[84,85] than with either ML or the models alone. We are not aware of any examples where models have been used in combination with ML for CATE estimation to improve forecasting accuracy compared with either alone and this is an important area for future research with regard to PK/PD, QSP, and disease progression models. This may be especially useful for those cases where there are relatively small clinical datasets, for example, those from some clinical trials.

POTENTIAL APPLICATIONS

Forecasting treatment effects at an individual level has great potential to transform the utility of observational data for clinical practice. Physicians currently rely on RCT evidence which, although it aims to handle unmeasured confounders through randomization, has limitations including HTEs, cost, and time required to perform them and limited generalizability of the evidence generated.

CATE estimation based on observational data could complement current RCT evidence in important ways. Current RCT evidence is routinely extrapolated to most patients who are excluded from them. However, accurate forecasting creates the opportunity to estimate which of those more complex patients would likely benefit from treatment. In this way, existing and new medicines could yield a greater overall individual and population-level health benefit. Currently, we expect that the clinical utility of a CATE-based treatment decision algorithm may need to be confirmed in an RCT. Such a trial would compare outcomes in a group of patients where the patient and physician have access to the forecasts from the CATE algorithm before making treatment choices, with a group where CATE estimates are not used. These trials would have as few inclusion/exclusion criteria as possible in order to ensure that they represent real-world patients.

CATE estimation could be useful for any disease where there are treatment choices, especially those with potentially serious consequences of failing to initiate effective treatment. Many cancers could benefit, where patients who will do poorly on first-line therapy could be started straight away on second or later lines of therapy. Rheumatoid arthritis is another example where CATE could help guide choosing which patients are best treated with anti-TNF drugs or anti-IL-6 to avoid structural joint damage developing or worsening, and when in their disease course to initiate them. Likewise, patients with severe asthma for which there are now several treatment options and where delays to initiating effective therapy could lead to patients suffering unnecessary exacerbations with a risk of death. CATE could also be useful to better understand the best treatment choices for populations, such as the elderly or those with multi-morbidities, who are frequently poorly represented in or even excluded from clinical trials.

Accurate forecasting also creates the potential to identify repurposable drugs where a drug developed for one indication has efficacy in an unrelated context. Hidden confounders are less likely with repurposing estimation as there is, by definition, no direct relationship between the treatment given and outcome observed. As repurposable drugs are typically already in routine clinical use, large observational datasets are a potentially rich source of training data for this indication.

CATE estimation could also be useful in drug development where clinical trial data could be used to develop an algorithm enabling estimation of CATE for individual patients receiving the drug being studied. CATE would complement the current statistical analysis of the effect in the population as a whole by providing a method to estimate individual treatment effects. If this could be done on the rather small datasets available in early development, the algorithm developed could be used to identify the patients to recruit into the phase III trial as those most likely to respond and/or to have the largest response. Alternatively, the CATE algorithm could be included in a phase III trial to confirm its clinical utility compared with treating patients without using information from the algorithm. Even if the phase III trial data were needed to develop an informative CATE algorithm, this would still provide a useful starting point for subsequent trials to demonstrate prospectively the usefulness of such an algorithm and allow the prescribing information to be updated. In any case, because of the previously mentioned limitation of RCTs to represent real-world patients, further development of the CATE algorithm will be required after the drug is on the market to refine it for use in the much wider patient population who will be likely to be treated with it. A general framework for this continuous refinement and updating has already been described.[86,87] Ultimately, in addition to statements of the population "average" benefit from a new treatment, a future product label for a drug with a CATE algorithm might indicate how to estimate the expected benefit in each individual patient or even require use of the algorithm in order to target the treatment to those most likely to have the best response.

Once there are enough clinical data to develop a CATE algorithm, it could be used to create virtual control populations to help streamline subsequent clinical trials. Indeed, a good CATE estimator has an advantage over population-based methods for creating virtual control populations as it offers the potential to forecast what the alternative treatment/control effect would have been in exactly the same patients as those receiving the test agent.

Accurate estimation of CATE could also uncover treatment efficacy that is otherwise obscured. For example, an RCT which does not meet its end point is considered "negative" and typically results in the treatment being discarded. However, although this means there was no observed difference in *average* group outcomes, there may be individuals that would yet benefit from treatment. Applying CATE analyses to RCT data could "salvage" treatments by identifying those who will respond and in whom treatment

should be considered.[50,88] However, application to RCT data in this way remains challenging as the approach would be subject to all the limitations of RCT described above (small size, restricted inclusion criteria minimizing heterogeneity, and therefore potentially treatable subgroups).

Finally, even when the end goal is not to make individualized inferences on treatment effects, but rather the study or discovery of more aggregate summaries of effect heterogeneity, such as subgroups or effect modifiers, strong ML-based CATE estimators could be of great use: they could be used to replace less sophisticated estimates that are currently used as inputs to a second stage for subgroup discovery[50,88] or for effect-modifier discovery.[89,90]

CHALLENGES AND LIMITATIONS

One of the most important limiting factors to producing high quality estimates of individualized treatment effects remains *the availability of good data*. On the one hand, domain expertise is needed to make a judgment whether all important variables are indeed recorded in the data, so that identifying assumptions are likely to hold. If important confounders are likely to be unmeasured, point estimation of effects is not possible unless further assumptions are made, and standard CATE estimators will output biased estimates. In such cases, it might be possible to produce instead a range of estimates for the CATE that are consistent with prespecified sensitivity models using recent advances in ML for creating prediction intervals under hidden confounding.[91–94] On the other hand, even if everything important is measured in the data, there also needs to be *enough of it*. Although it is not possible to make blanket assessments of exact sample sizes needed for different methods, the amount of data that are needed to produce good estimates generally increases with the number of covariates and the flexibility of the ML model. An often-cited rule of thumb, for example, suggests that sample sizes should be at least 10 times the number of free parameters in a model.[95] Most ML CATE estimators have nonetheless been developed and tested on datasets of moderate size: the most popular datasets IHDP and ACIC2016[82] have around 700 samples with 25 covariates and around 4,800 samples with 58 features, respectively. As systematic studies of sample size requirements have not yet been conducted for CATE estimation, future work may want to investigate this question in a manner similar to recent efforts in the context of clinical prediction models.[96]

Conversely, the less data are available, the more regularization (i.e., penalties for model complexity) is usually needed to avoid heavy overfitting on the training data. Thus, small data set sizes can clearly be at odds with some of the strategies discussed in this paper, such as the inclusion of higher-dimensional covariates to ensure unconfoundedness and the use of more flexible models to mitigate the effects of covariate shifts. Especially when one is doubtful whether the amount of data collected is sufficient for CATE estimates to be trustworthy, it can therefore be instructive to consult recent methods that have been developed to quantify their uncertainty and provide not only point estimates for the CATE but also confidence intervals[97–99]; if such intervals are very wide, this gives evidence that there are not enough data to have confidence in the individual predicted effects.

Another important challenge is the need to validate model outputs. This question is of course also deeply related to the discussion above -- model validation strategies are needed to choose between different classes of methods (and their hyperparameters) that differ in their implied flexibility. Two recent studies comparing different model selection criteria for this purpose can be found in references 65 and 66. Yet, even when a single best final model has been chosen, one may want to validate the obtained CATE estimates against available domain knowledge. One way of doing so is to make use of RCT evidence as the current gold standard by aggregating CATE estimates across individuals that meet the original trial's eligibility criteria and comparing this to the published trial results.[100] Another possibility would be to extract effect modifiers discovered by the models (for example, using variable importance measures as in refs. 89 and 90) and evaluate their plausibility using existing domain knowledge.

CONCLUSION

ML holds great potential for individualizing treatment effect estimation. However, in this context, much greater care is needed than in the standard prediction setting for which most ML methods were developed originally. Whereas ML expertise is needed to ensure smooth implementation of available methods, inclusion of domain expertise remains necessary to verify that untestable assumptions hold in applications. Further, adjustments for distribution shifts might need to be applied and the use of methods that were specifically *adapted* to target individual treatment effects instead of outcome prediction is recommended. It should also be recognized that, to be clinically useful, ML methods to estimate individual treatment effects do not need to be perfect but only to produce a clinically (and statistically) meaningful improvement in patient outcomes compared with using RCT data of average population level effect and assuming this correctly represents individual level effects in all real-world patients.

Validation of estimated individualized effects remains a challenge due to the fundamental problem of causal inference, and some or all of the strategies discussed in this paper might have to be applied to ensure the greatest level of trust in outputs. The lack of ground truth data for evaluation is also a problem during method development, which is why proof-of-concept for newly proposed methods is usually provided using simulated data that is often lacking grounding in real-world characteristics, and the ML community is missing good, realistic benchmark data sets to achieve meaningful progress on the treatment effect estimation task. There is thus an opportunity for the clinical pharmacology and other clinical communities to bring domain expertise and provide data sets that can act as better testbeds to enable development of even better methods, in particular those that are better able to deal with further complexities of real-world data.

FUNDING

A.C. is a PhD student funded by AstraZeneca.

CONFLICT OF INTEREST

R.W.P. receives compensation from and holds stock in F Hoffmann la Roche. J.W. receives compensation from and holds stock in AstraZeneca. The Cambridge Center for AI in Medicine which MvdS is

leading is funded by AstraZeneca and GSK. All other authors declare no conflicts of interest.

DISCLAIMER

As an Associate Editor of *Clinical Pharmacology & Therapeutics*, Richard W. Peck was not involved in the review or decision process for this paper.

1. Fortin, M., Dionne, J., Pinho, G., Gignac, J., Almirall, J. & Lapointe, L. Randomized controlled trials: do they have external validity for patients with multiple comorbidities? *Ann. Fam. Med.* **4**, 104–108 (2006).
2. Zulman, D.M., Sussman, J.B., Chen, X., Cigolle, C.T., Blaum, C.S. & Hayward, R.A. Examining the evidence: a systematic review of the inclusion and analysis of older adults in randomized controlled trials. *J. Gen. Intern. Med.* **26**, 783–790 (2011).
3. He, J., Morales, D.R. & Guthrie, B. Exclusion rates in randomized controlled trials of treatments for physical conditions: a systematic review. *Trials* **21**, 1–11 (2020).
4. Kennedy-Martin, T., Curtis, S., Faries, D., Robinson, S. & Johnston, J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials* **16**, 1–14 (2015).
5. Martin, K., Bégaud, B., Latry, P., Miremont-Salamé, G., Fourrier, A. & Moore, N. Differences between clinical trials and postmarketing use. *Br. J. Clin. Pharmacol.* **57**, 86–92 (2004).
6. Victora, C.G., Habicht, J.P. & Bryce, J. Evidence-based public health: moving beyond randomized trials. *Am. J. Public Health* **94**, 400–405 (2004).
7. Willke, R.J., Zheng, Z., Subedi, P., Althin, R. & Mullins, C.D. From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: a primer. *BMC Med. Res. Methodol.* **12**, 1–12 (2012).
8. Gabler, N.B., Duan, N., Liao, D., Elmore, J.G., Ganiats, T.G. & Kravitz, R.L. Dealing with heterogeneity of treatment effects: is the literature up to the challenge? *Trials* **10**, 1–12 (2009).
9. Angus, D.C. & Chang, C.C.H. Heterogeneity of treatment effect: estimating how the effects of interventions vary across individuals. *JAMA* **326**, 2312–2313 (2021).
10. Raman, G. *et al*. Evaluation of person-level heterogeneity of treatment effects in published multiperson N-of-1 studies: systematic review and reanalysis. *BMJ Open* **8**, e017641 (2018).
11. Kent, D.M. *et al*. The predictive approaches to treatment effect heterogeneity (PATH) statement. *Ann. Intern. Med.* **172**, 35–45 (2020).
12. Nestor, B. *et al*. Rethinking clinical prediction: why machine learning must consider year of care and feature aggregation (2018). *arXiv preprint arXiv:1811.12583*.
13. Personalized Medicine Coalition. Personalized Medicine at FDA. https://www.personalizedmedicinecoalition.org/Userfiles/PMC-Corporate/file/report.pdf. Accessed August 16, 2023. (2023).
14. Sparano, J.A. *et al*. Adjuvant chemotherapy guided by a 21-gene expression assay in breast. *Cancer* **379**, 111–121 (2018).
15. Fendt, R. *et al*. Data-driven personalization of a physiologically based pharmacokinetic model for caffeine: a systematic assessment. *CPT Pharmacometrics Syst. Pharmacol.* **10**, 782–793 (2022).
16. Mostafa, S. *et al*. Delineating gene–environment effects using virtual twins of patients treated with clozapine. *CPT Pharmacometrics Syst. Pharmacol.* **12**, 168–179 (2023).
17. Gaweda, A.E., Lederer, E.D. & Brier, M.E. Artificial intelligence-guided precision treatment of chronic kidney disease-mineral bone disorder. *CPT Pharmacometrics Syst. Pharmacol.* **10**, 1305–1315 (2022).
18. Bica, I., Alaa, A.M., Lambert, C. & Van Der Schaar, M. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clin. Pharmacol. Ther.* **109**, 87–100 (2021).
19. Rubin, D.B. Causal inference using potential outcomes: design, modeling, decisions. *J. Am. Stat. Assoc.* **100**, 322–331 (2005).
20. Holland, P.W. Statistics and causal inference. *J. Am. Stat. Assoc.* **81**, 945–960 (1986).
21. van der Laan, M.J. & Robins, J.M. *Unified Methods for Censored Longitudinal Data and Causality* (Springer, New York, 2003).
22. Rosenbaum, P.R. & Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).
23. Imbens, G.W. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev. Econ. Stat.* **86**, 4–29 (2004).
24. Vapnik, V. *The Nature of Statistical Learning Theory* (Springer Science & Business Media, New York, 1995).
25. Hastie, T., Tibshirani, R. & Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2 (Springer, New York, 2009).
26. Hernán, M. & Robins, J. *Causal Inference: What If* (Chapman & Hall/CRC, Boca Raton, 2020).
27. Schisterman, E.F., Cole, S.R. & Platt, R.W. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology* **20**, 488–495 (2009).
28. Cinelli, C., Forney, A. & Pearl, J. A crash course in good and bad controls. *Sociol. Method Res.* (2022). https://doi.org/10.1177/00491241221099552.
29. Zeng, J., Gensheimer, M.F., Rubin, D.L., Athey, S. & Shachter, R.D. Uncovering interpretable potential confounders in electronic medical records. *Nat. Commun.* **13**, 1014 (2022).
30. Louizos, C., Shalit, U., Mooij, J.M., Sontag, D., Zemel, R. & Welling, M. Causal effect inference with deep latent-variable models. *Adv. Neural Inf. Process. Syst.* **30** (2017).
31. D'Amour, A., Ding, P., Feller, A., Lei, L. & Sekhon, J. Overlap in observational studies with high-dimensional covariates. *J. Econom.* **221**, 644–654 (2021).
32. D'Amour, A. & Franks, A. Deconfounding scores: feature representations for causal effect estimation with weak overlap (2021). *arXiv preprint arXiv:2104.05762*.
33. Wu, P.A. & Fukumizu, K. $\beta $-intact-VAE: identifying and estimating causal effects under limited overlap. International Conference on Learning Representations (2022).
34. Gui, L. & Veitch, V. Causal estimation for text data with (apparent) overlap violations (2022). *arXiv preprint arXiv:2210.00079*.
35. Brookhart, M.A., Wyss, R., Layton, J.B. & Sturmer, T. Propensity score methods for confounding control in nonexperimental research. *Circ. Cardiovasc. Qual. Outcomes* **6**, 604–611 (2013).
36. Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Stat. Plann. Infer.* **90**, 227–244 (2000).
37. Wen, J., Yu, C.-N. & Greiner, R. Robust learning under uncertain test distributions: relating covariate shift to model misspecification. International Conference on Machine Learning, 631–639, PMLR. (2014).
38. Alaa, A. & van der Schaar, M. Limits of estimating heterogeneous treatment effects: guidelines for practical algorithm design (2018). International Conference on Machine Learning, 129–138, PMLR.
39. Farahani, A., Voghoei, S., Rasheed, K. & Arabnia, H.R. A brief review of domain adaptation. Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020, 877–894. (2021).

40. Ganin, Y. *et al*. Domain-adversarial training of neural networks. *J Mach Learn Res* **17**(59), 1–35 (2016).
41. Johansson, F., Shalit, U. & Sontag, D. Learning representations for counterfactual inference International Conference on Machine Learning, 3020–3029, PMLR. (2016).
42. Johansson, F.D., Kallus, N., Shalit, U. & Sontag, D. Learning weighted representations for generalization across designs (2018). *arXiv preprint arXiv:1802.08598*.
43. Shalit, U., Johansson, F.D. & Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. International Conference on Machine Learning, 3076–3085, PMLR. (2017).
44. Hassanpour, N. & Greiner, R. Counterfactual regression with importance sampling weights (2019a). In IJCAI, 5880–5887.
45. Hassanpour, N. & Greiner, R. Learning disentangled representations for counterfactual regression (2019b). International Conference on Learning Representations.
46. Assaad, S. *et al*. Counterfactual representation learning with balancing weights International Conference on Artificial Intelligence and Statistics, 1972–1980, PMLR. (2021).
47. Byrd, J. & Lipton, Z. What is the effect of importance weighting in deep learning? International Conference on Machine Learning, 872–881, PMLR. (2019).
48. Hendrycks, D. & Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations (2019). *arXiv preprint arXiv:1903.12261*.
49. Tripuraneni, N., Adlam, B. & Pennington, J. Overparameterization improves robustness to covariate shift in high dimensions. *Adv. Neural Inf. Process. Syst.* **34**, 13883–13897 (2021).
50. Foster, J.C., Taylor, J.M. & Ruberg, S.J. Subgroup identification from randomized clinical trial data. *Stat. Med.* **30**, 2867–2880 (2011).
51. Hill, J.L. Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Stat.* **20**, 217–240 (2011).
52. Ballman, K.V. Biomarker: predictive or prognostic? *J. Clin. Oncol.* **33**, 3968–3971 (2015).
53. Kunzel, 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. Van der Laan, M.J. & Rose, S. *Targeted Learning: Causal Inference for Observational and Experimental Data*, Vol. **4** (Springer, New York, 2011).
55. Athey, S., Tibshirani, J. & Wager, S. Generalized random forests. *Ann. Stat.* **47**, 1148–1178 (2019).
56. Nie, X. & Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108**, 299–319 (2021).
57. Kennedy, E.H. Optimal doubly robust estimation of heterogeneous causal effects (2020). *arXiv preprint arXiv:2004.14497*.
58. Curth, A. & van der Schaar, M. Nonparametric estimation of heterogeneous treatment effects: from theory to learning algorithms. International Conference on Artificial Intelligence and Statistics, 1810–1818, PMLR. (2021).
59. Imai, K. & Ratkovic, M. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann Appl Stat* **7**, 443–470 (2013).
60. Hahn, P.R., Murray, J.S. & Carvalho, C.M. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Anal.* **15**, 965–1056 (2020).
61. Curth, A. & van der Schaar, M. On inductive biases for heterogeneous treatment effect estimation. *Adv. Neural Inf. Process. Syst.* **34**, 15883–15894 (2021).
62. Rolling, C.A. & Yang, Y. Model selection for estimating treatment effects. *J. R. Stat. Soc. Series B Stat. Methodology* **76**, 749–769 (2014).
63. Schuler, A., Baiocchi, M., Tibshirani, R. & Shah, N. A comparison of methods for model selection when estimating individual treatment effects (2018). *arXiv preprint arXiv:1804.05146*.
64. Saito, Y. & Yasui, S. Counterfactual cross-validation: stable model selection procedure for causal inference models. International Conference on Machine Learning, 8398–8407, PMLR. (2020).
65. Mahajan, D., Mitliagkas, I., Neal, B. & Syrgkanis, V. Empirical analysis of model selection for heterogenous causal effect estimation (2022). *arXiv preprint arXiv:2211.01939*.
66. 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. International Conference on Machine Learning, PMLR. (2023).
67. Berrevoets, J., Imrie, F., Kyono, T., Jordon, J. & van der Schaar, M. To impute or not to impute? Missing data in treatment effect estimation. International Conference on Artificial Intelligence and Statistics, 3568–3590, PMLR. (2023).
68. Bica, I., Jordon, J. & van der Schaar, M. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Adv. Neural Inf. Process. Syst.* **33**, 16434–16445 (2020).
69. Schwab, P., Linhardt, L., Bauer, S., Buhmann, J.M. & Karlen, W. Learning counterfactual representations for estimating individual dose-response curves. *Proc. AAAI Conf. Artif. Intell.* **34**, 5612–5619 (2020).
70. Kaddour, J., Zhu, Y., Liu, Q., Kusner, M.J. & Silva, R. Causal effect inference for structured treatments. *Adv. Neural Inf. Process. Syst.* **34**, 24841–24854 (2021).
71. Kyono, T., Bica, I., Qian, Z. & van der Schaar, M. Selecting treatment effects models for domain adaptation using causal knowledge. *ACM Transact. Comput. Healthc.* **4**, 1–29 (2023).
72. Chapfuwa, P., Assaad, S., Zeng, S., Pencina, M.J., Carin, L. & Henao, R. Enabling counterfactual survival analysis with balanced representations. Proceedings of the Conference on Health, Inference, and Learning, 133–145. (2021).
73. Curth, A., Lee, C. & van der Schaar, M. Survite: learning heterogeneous treatment effects from time-to-event data. *Adv. Neural Inf. Process. Syst.* **34**, 26740–26753 (2021).
74. Curth, A. & van der Schaar, M. Understanding the impact of competing events on heterogeneous treatment effect estimation from time-to-event data. International Conference on Artificial Intelligence and Statistics, 7961–7980. PMLR. (2023).
75. Lim, B., Alaa, A. & van der Schaar, M. Forecasting treatment responses over time using recurrent marginal structural networks. *Adv. Neural Inf. Process. Syst.* **31** (2018).
76. 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 (2019).
77. Melnychuk, V., Frauen, D. & Feuerriegel, S. Causal transformer for estimating counterfactual outcomes. In Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, 15293–15329. PMLR (2022).
78. Seedat, N., Imrie, F., Bellot, A., Qian, Z. & van der Schaar, M. Continuous-time modeling of counterfactual outcomes using neural controlled differential equations International Conference on Machine Learning, 19497–19521, PMLR. (2022).
79. De Brouwer, E., Gonzalez, J. & Hyland, S. Predicting the impact of treatments over time with uncertainty aware neural differential equations. International Conference on Artificial Intelligence and Statistics, 4705–4722, PMLR (2022).
80. Vanderschueren, T., Curth, A., Verbeke, W. & van der Schaar, M. Accounting for informative sampling when learning to forecast treatment outcomes over time. International Conference on Machine Learning (ICML). PMLR (2023).
81. Dehghani, M. *et al*. The benchmark lottery (2021). *arXiv preprint arXiv:2107.07002*.
82. 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. Adv. Neural Inf. Process. Syst. 34 (2021).
83. Qian, Z., Zame, W., Fleuren, L., Elbers, P. & van der Schaar, M. Integrating expert ODEs into neural ODEs: pharmacology and disease progression. *Adv. Neural Inf. Process. Syst.* **34**, 11364–11383 (2021).

84. Yauney, G. & Shah, P. Reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection Machine Learning for Healthcare Conference, 161–226, PMLR. (2018).
85. Ribba, B., Bräm, D.S., Baverel, P.G. & Peck, R.W. Model enhanced reinforcement learning to enable precision dosing: a theoretical case study with dosing of propofol. *CPT Pharmacometrics Syst. Pharmacol.* **11**, 1497–1510 (2022).
86. Peck, R.W., Weiner, D., Cook, J. & Powell, J.R. A real-world evidence framework for optimizing dosing in all patients with COVID-19. *Clin. Pharmacol. Ther.* **108**, 921–923 (2020).
87. Powell, J.R., Cook, J., Wang, Y., Peck, R.W. & Weiner, D. Drug dosing recommendations for all patients: a roadmap for change. *Clin. Pharmacol. Ther.* **109**, 65–72 (2021).
88. Deng, C. *et al*. Practical guidance on modeling choices for the virtual twins method. *J. Biopharm. Stat.* **1–24**, 653–676 (2023).
89. Crabbé, J., Curth, A., Bica, I. & van der Schaar, M. Benchmarking heterogeneous treatment effect models through the lens of interpretability. *Adv. Neural Inf. Process. Syst.* **35**, 12295–12309 (2022).
90. Hermansson, E. & Svensson, D. On discovering treatment-effect modifiers using virtual twins and causal forest ML in the presence of prognostic biomarkers International Conference on Computational Science and Its Applications, 624–640. Springer International Publishing, Cham. (2021).
91. Jesson, A., Mindermann, S., Gal, Y. & Shalit, U. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. International Conference on Machine Learning, 4829–4838, PMLR (2021).
92. 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).
93. Oprescu, M., Dorn, J., Ghoummaid, M., Jesson, A., Kallus, N. & Shalit, U. B-learner: quasi-oracle bounds on heterogeneous causal effects under hidden confounding. International Conference on Machine Learning (ICML) (2023).
94. Yin, M., Shi, C., Wang, Y. & Blei, D.M. Conformal sensitivity analysis for individual treatment effects. *J. Am. Stat. Assoc.*, 1–14 (2022).
95. Baum, E. & Haussler, D. What size net gives valid generalization? *Adv. Neural Inf. Process. Syst.* **1** (1988).
96. Riley, R.D. *et al*. Calculating the sample size required for developing a clinical prediction model. *BMJ* **368** (2020). https://doi.org/10.1136/bmj.m441.
97. Jesson, A., Mindermann, S., Shalit, U. & Gal, Y. Identifying causal-effect inference failure with uncertainty-aware models. *Adv. Neural Inf. Process. Syst.* **33**, 11637–11649 (2020).
98. Lei, L. & Candès, E.J. Conformal inference of counterfactuals and individual treatment effects. *J. R. Stat. Soc. Series B Stat. Methodology* **83**, 911–938 (2021).
99. Lee, H.S., Zhang, Y., Zame, W., Shen, C., Lee, J.W. & van der Schaar, M. Robust recursive partitioning for heterogeneous treatment effects with uncertainty quantification. *Adv. Neural Inf. Process. Syst.* **33**, 2282–2292 (2020).
100. Qian, Z., Zhang, Y., Bica, I., Wood, A. & van der Schaar, M. Synctwin: treatment effect estimation with longitudinal outcomes. *Adv. Neural Inform. Process. Syst.* **34,** 3178–3190 (2021).