From prediction to prescription: Machine learning and Causal Inference

Judith Abécassis,[1] **Élise Dumas,**[2] **Julie Alberge,**[1] **and Gaël Varoquaux**[1,3]

[1]Soda, Inria Saclay, France; email: firstname.lastname@inria.fr

[2]Institute of Mathematics, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; email: elise.dumas@epfl.ch

[3]Probabl.ai, Paris, France

November 8, 2024

Abstract

The increasing accumulation of medical data brings the hope of data-driven medical decision-making, but its increasing complexity –as text or images in electronic health records– calls for complex models, such as machine learning. Here, we review how machine learning can be used to inform decisions for individualized interventions, a causal question. Going from prediction to causal effects is challenging as no individual is seen as both treated and not. We detail how some data can support some causal claims and how to build causal estimators with machine learning. Beyond variable selection to adjust for confounding bias, we cover the broader notions of study design that make or break causal inference. As the problems span across diverse scientific communities, we use didactic yet statistically precise formulations to bridge machine learning to epidemiology.

Keywords: personalized medicine, data-driven decision making, large scale data, causal inference, machine learning

Contents

1 Introduction: from prediction to individualized interventions

Data is getting richer and more complex: monitoring devices are cheaper, administrative records are ubiquitous, and notes are consolidated into databases. This trend brings the hope of better data-driven decision-making, such as individualized medicine. Machine learning (ML) tools are central to this hope, as they can build predictions from such complex data. Yet prediction often does not suffice: inferring causal effects is needed to decide on how to intervene. Causal effects characterize how an intervention in a system modifies an outcome of interest. Not all statistical associations are causal. For instance, a hospital visit is associated with an increase in mortality. This association results not from the hospital's effect but from the confounding effect of the worse baseline health of hospital-goers.

If the investigator can actually intervene, the straightforward approach to measuring causal effects is the randomized study: as the population with and without intervention are statistically alike, computing the average effect of the intervention does not require sophisticated statistics. But the average effect provides only a partial perspective: some individuals may respond to the intervention and others not. Personalized decision makings call for *individualized* or *conditional* causal effects, which do require more elaborate statistical approaches. Likewise, a randomized intervention is not possible in observational settings, causal inference must emulate randomized allocation, and statistical modeling is crucial. This review covers how machine learning can be used for this statistical modeling, opening the door to individualizing interventions thanks to causal inference from complex or high-dimensional data. Compared to existing reviews (Crown, 2019; Blakely et al., 2020; Prosperi et al., 2020; Curth et al., 2024; Liu, 2024), our focus is on an up-to-date, didactic but statistically precise, description of machine learning for estimating *heteregoneous* effects needed for personalized medicine.

The sophistication of modern machine-learning approaches and the amount of available data can capture subtle variation of effects effects across individuals. But, when working from observational data –without randomized intervention– a risk is always to capture non-causal associations resulting from current behaviors or policies. Causal validity requires a rigorous process. First, one must define causal questions that can be answered from the data at hand, as explained in Section 2. Then, different machine-learning approaches, detailed in Section 3, can then be used to individualize causal effects. But pitfalls in observational data can trick the most powerful machine learning approach, and Section 4 discusses how to avoid them with good study design.

2 Defining and identifying a causal effect

By themselves, machine learning methods provide only predictions. Causal interpretations, estimating the effects that ground data-driven decision-making, require more, namely a causal framework.

2.1 Potential outcomes and causal effects of interest

We use the formalization of causal effects from the Neyman-Rubin potential outcome framework (Imbens and Rubin, 2015). We consider the observed outcome $Y \in \mathbb{R}$, a function of an intervention or action $A \in \{0, 1\}$ –often called treatment or exposure. We use a binary treatment to illustrate the main notations and concepts. Still, the framework can be generalized to compare any two treatment levels –for discrete categorical and ordered or continuous treatments. The central idea of the potential outcomes formulation of causality is to consider for each individual the outcomes that would occur under any value of the intervention A. The two potential outcomes are then written $(Y(0), Y(1))$. Causal effects can be defined as quantities –called estimands– based on contrasts between the potential outcomes $(Y(0), Y(1))$. For personalized decision making, the best causal quantity would be the Individual Treatment Effect (ITE): $Y(1) - Y(0)$. However, the ITE can never be observed, as a subject has only one version of the treatment. This is the fundamental problem of causal inference (Holland, 1986). At the population level, some causal quantities can be accessible:

Definition 2.1 (Average treatment effect)**.** *The "Average Treatment Effect":*

ATE $$\tau = \mathbb{E}[Y(1) - Y(0)], \tag{1}$$

ATE Average Treatment Effect: the average difference between treated and not treated across the whole population

Computed across the whole population, the ATE summarizes the treatment effect which can be different for each individual. As it lumps together very different individuals, another interesting causal quantity (estimand) is to consider a conditional average, considering relevant individual characteristics as covariates $\mathbf{X}_{CATE} \in \mathbb{R}^{d_{CATE}}$, *eg* age, general health status, stage of diagnosis of the treated condition etc:

Definition 2.2 (Conditional average treatment effect)**.**

CATE
$$\tau(\mathbf{x}_{CATE}) = \mathbb{E}[Y(1) - Y(0)|\mathbf{X}_{CATE} = \mathbf{x}_{CATE}] \tag{2}$$

CATE Conditional Average Treatment Effect: the average difference between treated and not treated for individuals with characteristics $\mathbf{X}_{CATE} = \mathbf{x}_{CATE}$

The challenge of causal inference is that in practice we do not get to observe the potential outcomes: the observed data are $(Y, \mathbf{X}, A)$, with $\mathbf{X}$ consisting of all observed covariates for an individual, including the covariates $\mathbf{X}_{\mathbf{CATE}}$ on which one can condition on to define and identify the $CATE$ estimands, $\mathbf{X} \in \mathbb{R}^d$. The $CATE$ is crucial to making individualized decisions based on a specific patient's characteristics. The above estimands are expressed in terms of potential outcomes, *ie* unobserved data. In some settings, these quantities can be identified to statistical estimands of the available observed variables. In the following sections, we detail this identification starting from the "ideal" case of the randomized controlled trial (RCT), and then considering more complicated settings.

2.2 Apples-to-apples comparisons: why Randomized Controlled Trials (RCTs) work

In a RCT, the treatment is given at random. Consequently, treated and untreated populations are statistically equivalent, and the ATE can simply be computed with the difference of the observed outcome between the two populations.

Identification expressing the target causal quantity with observable variables; requires non-testable assumptions.

Doing a formal mathematical proof of the above intuition for identification reveals the required assumptions:

Assumption 2.3. Stable Unit Treatment Value Assumption (SUTVA) *or Consistency*

SUTVA Assumption:
$$Y(a) = Y \quad \textit{if } A = a$$

SUTVA states that the observed outcome corresponds to the potential outcome of the treatment actually taken by the unit. In particular, it implies that there is only one version of the treatment and that there is no interference between the units: the outcome of a unit does not depend on the treatment received by another unit. Therefore, SUTVA is broken in several common cases, such as the vaccination, as the outcome of an individual depends on their own vaccinal status, but also on the degree of vaccination in the surrounding population (herd immunity). SUTVA can also be broken by binarizing a continuous treatment (with several possible doses) or when the treatment is ill-defined. For instance, a binary variable indicating smoking or not is simplistic, as smoking duration and intensity may induce very different outcomes. The problem of defining a suitable treatment with respect to the SUTVA assumption (also called consistency) requires a particular attention, as detailed in section 4.2.

In an RCT, randomization of the treatment assignment enforces the two other assumptions important for causal inference: *unconfoundedness* and *overlap*.

Assumption 2.4. Ignorability *(also called unconfoundedness) – The treatment assignment is independent from the potential outcomes:*
$$(Y(0), Y(1)) \perp\!\!\!\perp A.$$

Assumption 2.5. Overlap *(also called positivity) – Any unit has a non-zero probability of receiving any version of the treatment:*
$$0 < P(A = 1) < 1.$$

With those assumptions, the causal estimand based on potential outcomes can be identified to a statistical estimand that depends only on observed quantities:

$$\begin{aligned} \mathbb{E}[Y(1)] &= \mathbb{E}[Y(1)|A = 1] && \text{from overlap and ignorability} \\ &= \mathbb{E}[Y|A = 1] && \text{from SUTVA} \end{aligned} \tag{3}$$

It follows similarly that $\mathbb{E}[Y(0)] = \mathbb{E}[Y|A = 0]$ and thus the ATE is identified as such:

$$ATE = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0]. \tag{4}$$

Likewise, the CATE can be identified to $\mathbb{E}[Y|A=1,\mathbf{X}_{CATE}] - \mathbb{E}[Y|A=0,\mathbf{X}_{CATE}]$ with conditional versions of ignorability and overlap. These equalities are important, because they identify a causal quantity, defined from the two potential outcomes impossible to observe, to quantities directly observable.

2.3 When the treatment is not random: a population-shift point of view

Populations are not comparable In *observational* data, where the treatment assignment does not result from a randomized design as in an RCT but from other factors, the treated and untreated populations are not equivalent, and the above assumptions do not hold. For instance, the decision to transfer or not patients to intensive care depends on the severity of their condition, the expected benefit (*eg* are they strong enough to endure this level of care), their age etc; so the population that is transferred differs from the non-transferred population and can not be directly compared, as shown in Figure 1. Formally, the *Overlap* and *Ignorability* assumptions do not hold.

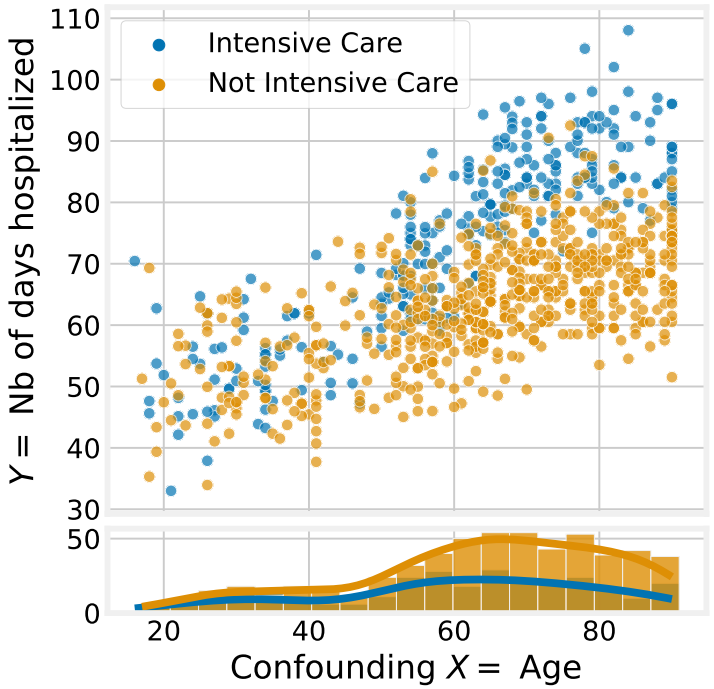

Figure 1: Example of observational data: The number of days a patient is hospitalized (Outcome) depends on whether they were transferred to intensive care (Treatment), and on the patient's age (Confounder). The patient's age influences also the likelihood of being transferred to intensive care. (Semi-simulated data)

Recovering identifiability: conditional ignorability and overlap Alternative and less stringent assumptions come into play: *conditional* versions of *Overlap* and *Ignorability*.

Assumption 2.6. Conditional ignorability *The treatment is independent of the potential outcomes, conditionally on a well-chosen set of covariates,* $\mathbf{X}_{ident} \in \mathbb{R}^{d_{ident}}$*:*

$$(Y(0), Y(1)) \perp\!\!\!\perp A \,|\, \mathbf{X}_{ident}.$$

Assumption 2.7. Conditional overlap *Any unit has a non-zero probability of receiving any version of the treatment, conditionally on covariates:*

$$0 < P(A = 1|\mathbf{X}_{ident} = \mathbf{x}_{ident}) < 1.$$

Here we denote the covariates necessary for the conditional overlap and ignorability assumptions to hold $\mathbf{X}_{ident}$ a different notation from the covariates used in the CATE (definition 2.2), but some covariates can belong to both sets: we assume that both covariate sets are observed, *ie* $\mathbf{X}_{ident} \cup \mathbf{X}_{CATE} \subset \mathbf{X}$. The

covariates used in the CATE are chosen to analyze treatment effect heterogeneity, ensuring conditional ignorability and overlap necessitates a meticulous selection of covariates, as elaborated in Section 2.4.

Intuitively, conditional overlap states that an individual has a comparable counterpart in the other treatment group for any possible set of characteristics, making it possible to compare treated and untreated individuals to estimate a causal effect. Conditional ignorability means that the treatment is as good as randomly assigned among the subjects with the same characteristics. In other words, with conditional versions of overlap and ignorability, among subjects with the same characteristics $\mathbf{x}_{ident}$, the data is similar to that obtained with an RCT, hence allowing to estimate a causal effect.

This intuition is maintained in the formal identifiability proofs: identification for the ATE proceeds through identification of the CATE for each possible value of $\mathbf{x}_{ident}$. Key to identification is that a potential outcome, such as $Y(1)$, can be written as an expectation of the observed outcomes Y reweighted by the probability of treatment:

$$
\begin{aligned}
\mathbb{E}[Y(1)] &= \mathbb{E}[\mathbb{E}[Y(1)|\mathbf{X}_{ident} = \mathbf{x}_{ident}]] && \text{from the law of total expectation} \\
&= \mathbb{E}[\mathbb{E}[Y(1)|A = 1, \mathbf{X}_{ident} = \mathbf{x}_{ident}]] && \text{from conditional ignorability and overlap} \\
&= \mathbb{E}[\mathbb{E}[Y|A = 1, \mathbf{X}_{ident} = \mathbf{x}_{ident}]] && \text{from SUTVA} \qquad (5) \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{Y\,\mathbb{1}_{A=1}}{\mathbb{P}(A = 1|\mathbf{X}_{ident} = \mathbf{x}_{ident})}|\mathbf{X}_{ident} = \mathbf{x}_{ident}\right]\right] \\
&= \mathbb{E}\left[\frac{Y\,\mathbb{1}_{A=1}}{\mathbb{P}(A = 1|\mathbf{X}_{ident} = \mathbf{x}_{ident})}\right] && \qquad (6)
\end{aligned}
$$

Strategies to estimate causal estimands Equations 5 and 6 form the basics of estimation strategies for causal estimands from observations:

Outcome prediction Equation 5 suggests modeling the conditional expectation of outcomes given the covariates and using their difference to obtain the causal effect. This approach makes a link to the intuitive causal reasoning behind mechanistic models, where changes in the input induce changes in the output in a causal way. Note that this causal interpretation holds for predictive models only if the SUTVA, conditional overlap, and conditional ignorability assumptions hold (which requires a good choice of covariates) and the conditional expectations in eq 5 are well estimated.

Inverse propensity weighting (IPW) Alternatively, equation 6 grounds the so-called IPW approaches based on reweighing observations. The weights –the inverse of propensity scores– depend on the values of the treatment and the covariates $\mathbf{X}_{ident}$: $w = 1/\mathbb{P}(A = 1|\mathbf{X}_{ident} = \mathbf{x}_{ident})$. This reweighing makes the covariate distributions comparable across the treated and untreated populations. The propensity score "summarizes" the influence of the (high-dimensional) covariates into a single number so that it can cancel out.

2.4 Causal graphs to explicit prior knowledge and causal assumptions

We have just seen that the identifiability assumptions, SUTVA, conditional ignorability (2.6), and overlap (2.7) are crucial to estimate causal effects. Conditional ignorability needs a good choice of covariates $\mathbf{X}_{ident}$. We now describe a framework useful for this choice, causal graphs (Greenland and Brumback, 2002; Tennant et al., 2021).

Confounding bias and spurious causal associations The prototypical threat to ignorability is confusion bias, where a third (confounding) variable explains the seemingly causal association between the treatment A and the outcome Y seen in the data. Failure to adjust for confounding bias can lead to a phenomenon known as Simpson's paradox (Simpson, 1951), which occurs when a trend between the treatment and the outcome in the whole dataset is different when considering subgroups defined by a third (confounding) variable. Simpson's paradox was noted in comparing treatments A and B for kidney stones (Charig et al., 1986; Julious and Mullee, 1994). Treatment A performed better for both small and large stones individually, but when data were combined, treatment B seemed superior. This paradox occurs because treatment A was mainly used for large stones, which have lower success rates. Without accounting for stone size, treatment A appeared less effective overall.

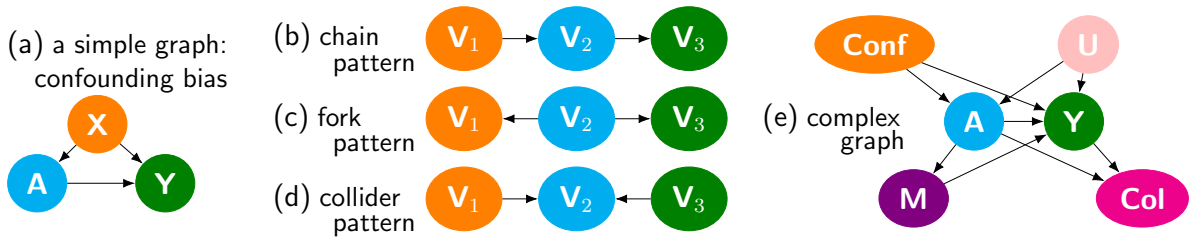

Figure 2: Causal graphs: (a) A simple causal graph, the treatment A has a causal effect on the outcome of interest Y; but a third variable X influences but A and Y.(b) A chain pattern, (c) a fork pattern, (d) a collider pattern, (e) a more complex causal graph, where "Conf" stands for confounder, "U" unobserved, "M" mediator, and "Col" collider.

One might think that including many variables in the confounding set, $\mathbf{X}_{ident}$, is a good idea to avoid confounding bias, making the ignorability assumption more plausible. However, by doing so, we may introduce other biases (MacKinnon and Lamp, 2021; Schisterman et al., 2009; Cinelli et al., 2024). A famous example of bias induced by over-adjustment is known as Berkson's paradox. Berkson's paradox manifests in a very similar way to Simpson's paradox, with a different correlation trend in the whole population or a subset, but the underlying causal mechanism is different: if we study individuals selected because they show degraded health –for instance hospitalized– we will find anti-correlations between causes: if a patient was not hospitalized for a stroke, he must have had another reason (Berkson, 1946). Such selection bias breaks causal inference and keeps happening, for instance, when investigating COVID-19 by focusing on symptomatic individuals (Griffith et al., 2020). The use of a causal graph can provide clear guidance to select the right adjustment set and establish whether conditional ignorability holds for a given causal question while avoiding the caveats of over-adjustment.

What is a causal graph? A causal graph describes the causal links between the variables of the problem, including unobserved variables. The corresponding variables are represented as nodes connected by directed edges (arrows). A causal graph is defined as a *directed acyclic graph* (DAG): all edges are directed and form no cycles. Figure 2a shows the confounding bias described above written as a (very simple) causal graph: the treatment A, the outcome of interest Y, and a third variable X. Arrows between them represent their causal relationships: the treatment A has a causal effect on the outcome Y, but the variable X, the stone size, also influences both A and Y. X is called a *confounder*. Intuitively, a confounder explains part of the association between A and Y. This part of the $A-Y$ association is thus not causal, and to isolate the causal effect, we need to correct the non-causal effect due to X. In that simple case, adjusting for X suffices to obtain identifiability using sec. 2.3. As we will see, using a graph enables to go from domain assumptions to the choice of the right variables in $\mathbf{X}_{ident}$.

Confounder A variable influencing both treatment assignment and outcome. Unaccounted for, it creates a non-causal association between treatment and outcome.

Identifiability criterion using the graph To formalize and generalize the idea of a causal and a non-causal part of the association between two variables, we can introduce the concept of *path*. A path in the graph is a sequence of at least two different nodes, such that there is an edge between each node and its successor in the sequence, regardless of its direction. If a path exists between A and Y, with all the edges oriented in the same direction, we say that this path is causal. In the graph of Figure 2a, there are two paths from A to Y: a causal path $A \rightarrow Y$, and a non-causal path $A \leftarrow X \rightarrow Y$. If a node in the path is the target of two distinct edges in the path, it is called a *collider relative to this path*. In the graph of Figure 2d, V_2 is a collider on the path $V_1 \rightarrow V_2 \leftarrow V_3$.

Paths can be opened or blocked. If we do not condition on any variable, a path is open unless there is a collider on it. If we condition on a non-collider node on a path, it blocks the path, but if we condition on a collider node, it opens the path. Overall, a path is blocked if there is at least one "blocking node" on it. Intuitively, if the path from A to Y is open, information can be exchanged between those two variables through this path, *ie* the two variables will be non-independent in the data.

We can now establish a criterion for choosing which variables to adjust for in order to enforce the conditional ignorability hypothesis. We can identify the causal effect between an exposure A and an outcome Y if we can adjust on a subset of variables $\mathbf{X}_{ident}$ such that they block all the non-causal paths from A to Y, and leave open all the causal ones. The procedure of finding an adjustment subset can be automatized, for instance, in the software Dagitty (Textor et al., 2016). Satisfying conditional ignorability is only possible if the graph structure and observed variables allow it.

How to obtain a causal graph? The causal structure encoded in the graph enables precise identification conditions. However, it requires knowing this graph. There are two main approaches to obtaining a causal graph. The first is expert knowledge, where the graph is built by hand, ideally with the assistance of one or several experts of the application domain. The second is using *causal discovery* to construct the graph using available data (Peters et al., 2017; Huber, 2024). In both cases, there can be uncertainty in the graph structure, and therefore it might be relevant to assess the robustness of the causal effect estimation to variations in the graph.

When it comes to variable selection, the perfect is the enemy of the good The criterion for conditional ignorability obtained from the causal graph replaces classical heuristics to choose the variables to include in the adjustment set. Simple heuristics can indeed lead to invalid analysis. A first misguided strategy would be to adjust on all the available variables, to make sure that all confounders are accounted for. The main risk is to select variables that should not be adjusted for, such as colliders (node Col in Figure 2d), common effects of the treatment and the outcome (MacKinnon and Lamp, 2021), or mediators (node M in Figure 2d) that could block a causal path and bias the estimation of the causal effect. Adjusting for those nodes will introduce additional bias in the estimation instead of correcting for confounding bias. An alternative way to express this is to state that one should refrain from adjusting for post-treatment variables, *ie*, variables that are causally influenced by the treatment. A second misguided strategy is to selection variables by relying on correlations in the data instead of knowledge of the graph structure. This strategy may also lead to selecting colliders or mediators. Another strategy, suboptimal though not invalid, is to consider variables that are causes of both the action A and the outcome Y. This set selection might lead to overadjustment in the sense that a more parsimonious subset of variables would lead to a valid causal result. Reducing the number of variables for in the model is useful because it facilitates estimation (reducing the variance). Finally, adjustment sets with more confounders make the conditional ignorability assumption more plausible but, at the same time, reduces overlap (D'Amour et al., 2021). In some cases, it can be interesting to consider not adjusting on weak confounders to preserve a stronger overlap, which could result in a smaller overall bias, though this is difficult to assess in practice.

3 Estimation with machine learning

Once the target estimand is defined with the covariates needed for identification from the available data, the next step is to use this data to actually estimate the causal quantity of interest. Recent machine-learning methods bring new flexibility to such estimation beyond simple models traditionally used –typically linear models. This flexibility is particularly welcomed to handle larger and more complex datasets as with electronic health records. We will first outline the two main strategies for estimation, introduced in the previous section: sample reweighting and outcome modeling, and then more advanced methods that best use machine learning for CATE estimation. In each case, we will outline practical considerations for the implementation of those approaches.

3.1 Addressing the distribution shift: Inverse propensity scores

One of the challenges in estimating causal effects is the distribution shift between the two treatment groups (visible on fig. 1). Reweighting the observations can address this distribution shift (Dockès et al., 2021), and gives a basic principle to estimate causal effects, following equation 6. Estimation proceeds in two steps: *1)* first estimate the propensity scores (Rosenbaum and Rubin, 1983), which are the probability of receiving the intervention given the confounding covariates:

1) estimate propensity score: $$e(\mathbf{x}_{ident}) \stackrel{\text{def}}{=} \mathbb{P}(A = 1|\mathbf{X}_{ident} = \mathbf{x}_{ident}). \tag{7}$$

A machine-learning classifier can be used to estimate these probabilities. $e(\mathbf{x})$ is called a nuisance model as it is not the final objective of the estimation procedure. *2)* These probabilities are plugged in as inverse propensity scores to compute a re-weighted difference in means, an estimator of the ATE τ:

2) plug-in reweighed differences $$\widehat{\tau}_{\text{IPW}} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} Y_i \left(\frac{A_i}{\widehat{e}(\mathbf{X}_{ident_i})} - \frac{1 - A_i}{1 - \widehat{e}(\mathbf{X}_{ident_i})} \right) \tag{8}$$

Individualized decision making calls for estimating the dependency of the effect on covariates, the CATE $\tau(\mathbf{x})$. This estimation can be obtained by replacing the empirical mean in equation 8 by a regression of X on reweighed targets (Wager and Athey, 2018) –or pseudo outcomes– $Y_{IPW} \overset{\text{def}}{=} Y(\frac{A}{\widehat{e}(X)} - \frac{1-A}{1-\widehat{e}(X)})$. In practice, any machine-learning regressor can be used, learning to predict Y_{IPW} from $\mathbf{X}_{CATE}$.

These formulas rely on inverting estimated probabilities, and this inversion will amplify estimation noise in $\widehat{e}$ when it is close to 0 or 1. But these "extreme" propensity scores signify that for some observations, the treatment is almost deterministic, thus violating overlap (assumption 2.5). In such regions of the covariate space, the treatment may be very unlikely, maybe infeasible, and thus, it may be preferable to avoid having causal claims (Li et al., 2019; Oberst et al., 2020). One solution is "trimming", excluding observations with $e \notin [\alpha, 1-\alpha]$ for a choice of α (Crump et al., 2009). This approach, however, does not define explicitly the covariate space on which the causal estimates are valid, and a more explicit solution may be preferable (Oberst et al., 2020).

Estimating good propensity scores $\widehat{e}$ can be estimated with any classifier predicting A from $\mathbf{X}_{ident}$. However, here the goal is not good classification, but accurate probabilities. This goal requires selecting models not to maximize accuracy or area under the ROC curve –Receiver Operating Characteristic (Varoquaux and Colliot, 2023)– but using a *strictly proper scoring rule* (Gneiting and Raftery, 2007), such as log-loss or Brier score. A common misconception is that it suffices to measure and correct the *calibration error*. The calibration error measures whether a probabilistic classifier is overconfident or underconfident, and simple *recalibration* methods can correct it: *eg* isotonic recalibration (Niculescu-Mizil and Caruana, 2005) temperature scaling (Guo et al., 2017). Yet, a classifier with zero calibration error may be far from the conditional probability e (Perez-Lebel et al., 2022). As machine-learning models can be systematically over or under-confident (Guo et al., 2017; Minderer et al., 2021), recalibration techniques may be useful. But, all in all, proper scoring rules must drive every aspect of model selection, including confirming the benefit of recalibration.

3.2 Outcome models: predicting the unobserved counter-factual outcome

An approach to estimating the causal effect of a treatment that leads naturally to CATE estimates is by predicting the expected potential outcome with and without the treatment. As outlined in equation 5, these expected potential outcomes can be written as function of $\mathbb{E}[Y|A, \mathbf{X}]$. For this, *outcome modeling* also proceeds in two steps: *1)* first it estimates the *response function*,

$$\text{1) estimate response function:} \qquad \mu(a, \mathbf{x}) \overset{\text{def}}{=} \mathbb{E}[Y|A = a, \mathbf{X} = \mathbf{x}]. \tag{9}$$

We use the notation $\mathbf{X} = \mathbf{X}_{ident} \cup \mathbf{X}_{CATE}$, to simplify the notations and as we aim at estimating the CATE. $\widehat{\mu}$ is estimated by fitting a regression model of the outcome given the treatment and the covariates, *eg* with a base machine-learning model. 2) Then the estimate of the CATE is given by contrasting the predictions with two different treatment options:

$$\text{2) plug-in g-formula:} \qquad \widehat{\tau}_S(\mathbf{x}) = \widehat{\mu}(1, \mathbf{x}) - \widehat{\mu}(0, \mathbf{x}) \tag{10}$$

If a *single* predictive model is used for the regression in equation 9, this approach is called S-learner. It may be beneficial to fit *two* distinct models for the treatment and the control, an approach called T-learner (Künzel et al., 2019):

$$\widehat{\tau}_T(\mathbf{x}) = \widehat{\mu}_1(\mathbf{x}) - \widehat{\mu}_0(\mathbf{x}) \quad \text{with} \quad \mu_a(\mathbf{x}) \overset{\text{def}}{=} \mathbb{E}[Y|A = a, \mathbf{X} = \mathbf{x}] \tag{11}$$

Both estimators are unbiased if the base regression models are unbiased[1]. The S and the T-learner correspond to two different inductive biases: the patterns that they easily capture differ (Curth and Van der Schaar, 2021). The S-learner will foster similar response surfaces for the treated and control populations. In the simple case of using a linear model for $\widehat{\mu}$, if no interaction term between A and $\mathbf{X}$ is added, the model imposes the same slope for treated and non-treated, capturing a constant –and not heterogeneous– effect (fig. 3a). In the T-learner, the two response functions are not biased to resemble each other, as they are fitted separately on distinct subparts of the data, which comes at the cost of using less data. Which one to prefer? The S-learner is preferable if there is enough shared structure

Inductive bias How a given machine learning method favours certain patterns over others to generalize. It can be explicitly visible in the model form (*eg* linear), or implicit.

[1] To be precise, the no-bias results are asymptotic, *ie* characterize consistency.

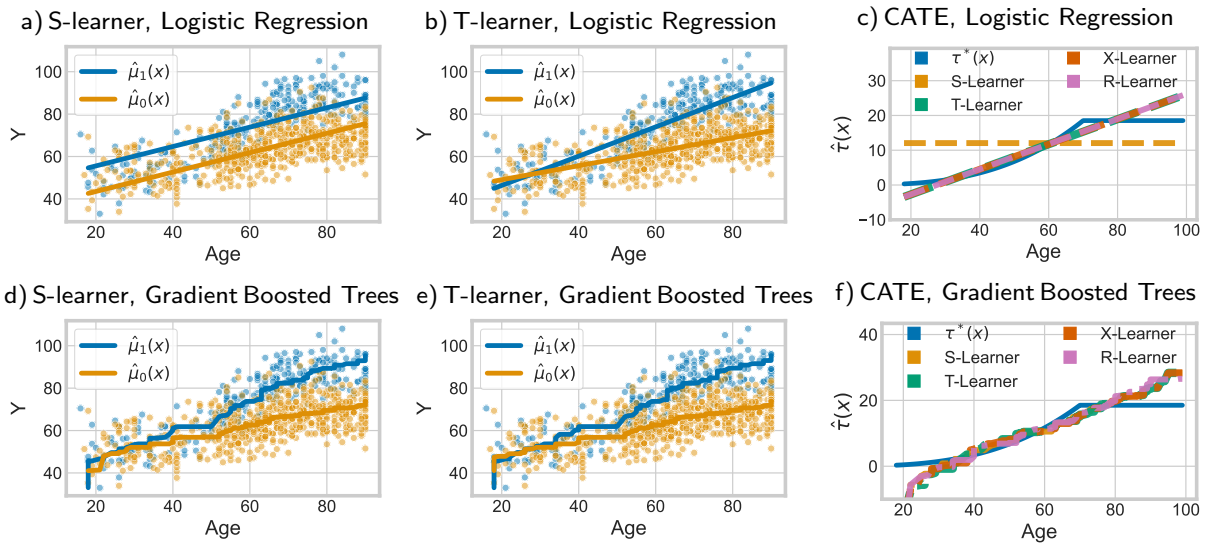

Figure 3: Different meta learners and base machine-learning algorithms. The upper plots display the estimations using parametric Linear Regression, including the response functions for the S-learner (a), the T-learner (b), and the CATE estimates for all meta-learners (c). The lower plots present the same estimations, but using Gradient Boosting Trees (GBT) for the estimation of nuisance parameters, including the S-learner (d), the T-learner (e), and the CATE estimates for all meta-learners (f).

between $\mu_0(\mathbf{x})$ and $\mu_1(\mathbf{x})$ to simplify markedly their estimation, a common setting in medicine (Hahn et al., 2020; Curth and Van der Schaar, 2021). Gauging this in practice is difficult, not only because one does not know beforehand μ_0 and μ_1, but also because the relevant notions of simplicity relate to the base machine-learning model used, with their own inductive biases. For a very flexible base model, there can little benefit of the T-learner compared to the S-learner (fig. 3d and 3e). Indeed, in regions where the two potential outcomes differ, a flexible S-learner will model these differently. Consider for instance a tree-based model (random forest, gradient-boosted trees), if the treatment is very predictive of the outcome, the first split of the trees will split on the treatment, thus the two population will subsequently be fitted separately.

The dilemma between the S and the T learner illustrates another challenge of estimating heterogeneous causal effects: while we can easily control the error on the response functions separately $\mu_0(\mathbf{x})$ and $\mu_1(\mathbf{x})$, we would like to minimize errors on $\tau(\mathbf{x}) = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x})$, a difference that we never observe, as no individual is simultaneously treated and not treated.

3.3 Imbalance between treatment groups and the X-learner

Often, there are much more observations in one treatment group than in the other. Typically, there are much fewer treated individuals, *eg* in administrative databases. A variant of this problem is the distribution shift across treatment groups, which creates an imbalance for some regions of the covariate space, with very few healthy people receiving a given treatment. The challenge is that the response function of the treated $\mu_1(\mathbf{x})$ is harder to estimate; it will come with more error than $\mu_0(\mathbf{x})$, either in the form of variance or in the form of an oversimplified function as the corresponding machine-learning model is more regularized. If there is a lot of shared structure between the two treatment groups, the S-learner can help, but still faces the conundrum of choosing the complexity of the function to both groups adequately, with widely differing data points.

The X-learner tackles this challenge by bringing information from one group to the other (Künzel et al., 2019). It first estimates the two response functions as the T-learner. Then, for each treatment group, it uses them to compute individual treatment effects. For instance, for a treated individual i, it infers the treatment effect $D(1)_i$ computing the difference between the observed outcome Y_i and the potential outcome as predicted by $\widehat{\mu}_0$: $D(1)_i = Y_i - \widehat{\mu}_0(\mathbf{X}_i)$). Importantly, this formula combines information from both groups, as the control response function $\widehat{\mu}_0$ is used on the treated group. A first machine-learning estimate of the CATE, $\widehat{\tau}_1$ is then computed by fitting $\widehat{D}(1)$ as a function of $\mathbf{X}$. A second estimate $\widehat{\tau}_0$

is computed by reversing the role of treated and controls. The final estimate is obtained by a weighted combination of estimates on both groups: $\widehat{\tau}(\mathbf{x}) = \widehat{e}(\mathbf{x})\,\widehat{\tau}_0(\mathbf{x}) + (1-\widehat{e}(\mathbf{x}))\,\widehat{\tau}_1(\mathbf{x})$. A key aspect here is that the weights $\widehat{e}$ favor the relative estimates where they are more trustworthy: where the treatment is likely, e is high, putting more weights on $\widehat{\tau}_0$, $\widehat{\tau}_0$ which is itself estimated via $\widehat{\mu}_1$ and therefore less noisy where there are many treated units.

3.4 Combining outcome models and propensity scores

Departing from outcome models, capture the link between $\mathbf{X}$ and Y, IPW methods model the probability of being treated, the link between $\mathbf{X}$ and A, to address the distribution shift between treatment groups. The two families of approaches, based on IPW or outcome models, lead to different errors, and it is useful to combine them.

R-decomposition and R-loss As one of the challenges is to separate out the treatment effect $\tau(\mathbf{x})$, the difference between the two treatment groups, from the common effects, the variations of baseline risk, it is useful to rewrite the problem introducing the mean outcome:

Conditional mean outcome: $$m(\mathbf{x}) \stackrel{\text{def}}{=} \mathbb{E}[Y|\mathbf{X}=\mathbf{x}] \tag{12}$$

The outcome can then be written (Robinson, 1988; Nie and Wager, 2021):

R-decomposition: $$Y(A) = m(\mathbf{X}) + \big(A - e(\mathbf{X})\big)\,\tau(\mathbf{X}) + \varepsilon(\mathbf{X}, A), \tag{13}$$

where, importantly, $\mathbb{E}[\varepsilon(\mathbf{X}, A)|\mathbf{X}, A] = 0$. This rewriting shows that, given estimates of m and e, τ can be readily estimated from the data by optimizing $\widehat{\tau}$ to minimize ε. Specifically, it suggests a risk to minimize (Nie and Wager, 2021; Chernozhukov et al., 2018; Foster and Syrgkanis, 2023; van der Laan and Luedtke, 2014):

R-Risk $$R - risk(\tau_f) \stackrel{\text{def}}{=} \mathbb{E}\big[(Y - m(\mathbf{X}) - (A - e(\mathbf{X}))\tau_f(\mathbf{X}))^2\big] \tag{14}$$

where τ_f is the candidate CATE that is optimized. The CATE $\widehat{\tau}(\mathbf{x})$ can then be estimated by fitting a machine-learning model using as a loss the above R-risk. This approach can easily be implemented even with machine-learning toolkits that do not support custom losses but do support sample weights. Any regression model (based on a squared loss) can be adapted by fitting pseudo outcomes[2] $Y_R \stackrel{\text{def}}{=} \frac{Y - \widehat{m}(\mathbf{X})}{A - \widehat{e}(\mathbf{X})}$ and using as sample weights[3] $W_R \stackrel{\text{def}}{=} (A - \widehat{e}(\mathbf{X}))^2$.

Augmented IPW and DR-learner Another route that leads to a risk combining outcome model and IPW is to consider corrections of the individual methods. Indeed, as discussed previously, in outcome modeling, the estimations of μ do not minimize the error on τ. The theory of influence functions can give corrections (Robins and Rotnitzky, 1995; Hahn, 1998), leading to define a pseudo-outcome, known as AIPW:

$$Y_{AIPW} \stackrel{\text{def}}{=} \mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) + \frac{A}{e(\mathbf{X})}\big(Y - \mu_1(\mathbf{X})\big) + \frac{1-A}{1-e(\mathbf{X})}\big(Y - \mu_0(\mathbf{X})\big) \tag{15}$$

Y_{AIPW} can be seen as providing corrections to a simple outcome-model estimator of $\tau(\mathbf{x})$: $\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})$. The corresponding correction is an IPW applied to the *residuals* of the outcome-model estimate. The formula can be rewritten to expose another, symmetric, interpretation:

$$Y_{AIPW} = Y\left(\frac{A}{e(\mathbf{X})} - \frac{1-A}{1-e(\mathbf{X})}\right) + \frac{e(\mathbf{X}) - A}{e(\mathbf{X})}\mu_1(\mathbf{X}) + \frac{1 - e(\mathbf{X}) - (1-A)}{1-e(\mathbf{X})}\mu_0(\mathbf{X}) \tag{16}$$

Here, Y_{AIPW} appears that an IPW estimate (eq. 8), with a correction that corresponds to an outcome model reweighed by the residual treatment probabilities.

A CATE estimate, called DR-learner, can be built by using a machine-learning model to regress Y_{AIPW} on $\mathbf{X}$ (Kennedy, 2023). The DR-learner divides by propensity scores, which will create noise for regions of extreme propensity scores (low overlap), unlike the R-learner which ignores the corresponding samples.

[2] for numerical stability, it can be useful to add ϵ, typically 10^{-6}, to the denominator. Note that not adding this ϵ to the sample weights W_R will shrink to zero samples with no reasonable matching counterparts, *ie* in regions without overlap, thus letting the inductive bias of the model form (trees, linear model) fill in.

[3] Without the sample weights, the regression would give a U-learner (Künzel et al., 2019; Nie and Wager, 2021), which is unstable with extreme propensity scores

Cross-fitting The above formulas (eq. 14 and 15) need estimates of nuisances e, μ, m, which must be computed beforehand, using a machine-learning model. To avoid coupling of estimation error of this first estimation in the second estimation that regresses on Y_R or Y_{AIPW}, the two steps should ideally be carried out on different samples. One approach is to split the data in two folds, fit the nuisance models on the first half, the CATE estimators on the second half, repeat the procedure swapping the two folds, and average the resulting CATE predictors.

Doubly robust property These CATE estimators perform well because the IPW errors (on $\widehat{e}$) and the outcome-modeling errors (on $\widehat{\mu}$ or $\widehat{m}$) cancel out. Asymptotically, only one of the two models (IPW or outcome model) need to be unbiased (well specified), for the CATE estimator to be unbiased (Bang and Robins, 2005): the corrections in equations 15 or 16 will be null. In finite samples, even unbiased estimators of IPW and outcome models will have estimation noise. However, in the CATE estimators, their errors are multiplied with one another, and as a result, they partly cancel out, giving fast convergence rates: fewer samples are needed to obtain good estimates (Nie and Wager, 2021; Chernozhukov et al., 2018; Kennedy et al., 2024).

3.5 Which base machine-learning model to use

The CATE estimators that we have covered here rely on regression models. As the specific functional form does not matter, any model can do, provided that it gives good estimates, *ie* predicts well. This opens the door to using models adapted to specific data, such as convolution neural networks for images or transformers for text. Text and images can be interesting in bringing more context to the individuals and thus capture potential confounding. However, as they typically mix information, it can be important to check that they do not contain post-treatment information, as colliders or mediators (MacKinnon and Lamp, 2021; Louizos et al., 2017; Veitch et al., 2020).

Super learners In causal-inference settings, *eg* health, the data are often of a tabular nature, on which tree models (such as gradient boosting) tend to work well (Grinsztajn et al., 2022). However, this is not a general rule; typically for low amount of data or or high noise settings linear models may be preferable. A popular approach in causal inference is the super-learner (Van der Laan et al., 2007), which can be implemented by using model stacking (Breiman, 1996) on a few complementary models, typical linear and tree-based models –modern autoML relies on a related notion of model portfolio (Feurer et al., 2022). A recent thorough benchmark of machine learning for causal inference confirmed that such stacking of linear model and gradient-boosted trees was a good solution for nuisance models (Doutreligne and Varoquaux, 2023).

Causal model selection Nuisance models can be selected by cross-validating in a standard machine-learning way. However, when selecting a model for the CATE it is important to keep in mind that the best model is not the best predictor in the usual sense (squared loss). Rather, a good approach is to measure the R-Risk (eq. 14) in a cross-validation loop (Doutreligne and Varoquaux, 2023). For this, nuisance models must be fitted in the train set in parallel to the CATE model, as they are needed for the R-Risk.

Different plausible machine-learning approaches can give different causal estimates (Bouvier et al., 2024; Doutreligne et al., 2023). An empirical comparison of method can assess the robustness of the result and grasp the key factors influencing the result. Such an approach, sometimes called vibration analysis (Patel et al., 2015; Doutreligne et al., 2023) can show how much the estimated effect varies across methods and which methods achieve similar or different performances.

4 No magic bullet for causal validity

Using data to personalize clinical decisions requires causal inference for each possible intervention, each with a tailored choice of causal estimand as well as an appropriate identification strategy (Hernán et al., 2019), multiple steps that require care and background knowledge (Hoffman et al., 2024). Indeed, unlike with classic prediction scenarios, as in machine-learning, we cannot just compare predicted values to observed values: we have never observed the same individual bother treated and not. Causal inference can go wrong in many ways. For instance, a flawed *study design* (the formulation of the question and

data to include) can lead to studying an impossible scenario such as going back in time and applying a prevention strategy to save an individual. These problems cannot be solved by accumulating large sample sizes or deploying more sophisticated procedures, *eg* based on machine-learning: if you aim at the wrong target, you'll miss, no matter how good an archer you are.

Study design overall strategy to answer a question, encompassing all the choices made for analyzing the collected data.

Let us consider a historical example: aspirin use and breast cancer recurrence. Several observational studies suggested that aspirin use in breast-cancer patients may reduce the risk of breast cancer recurrence (Chen and Holmes, 2017). As a result, aspirin was proposed as a potential treatment for breast cancer, and a large double-blind randomized trial was conducted, in which patients were randomized to receive either 300 mg of aspirin or a placebo daily for five years (Chen et al., 2024). Contrary to previous observational findings, the trial results were negative, showing no reduction in breast cancer recurrence risk with aspirin use. We use this example below to illustrate some points of attention when designing and running a causal analysis. However, we do not intend to provide a definitive explanation for the apparently different results between the randomized trial and the previous observational studies in the case of aspirin and breast cancer.

We first explore potential pitfalls resulting from ill-defined causal questions and flawed study designs.

4.1 On an Unambiguous Specification of the Causal Question

To know exactly where and when a causal inference is applicable to personalize decision, it is important to carefully specify what question it answer. The effect of an intervention is studied via a *causal estimand*: the contrast (*e.g.*, difference or ratio) between counterfactual outcomes under different interventions (*e.g.*, treatment versus control) within a population (Hernán and Robins, 2020). Several frameworks have been proposed to explicit a causal estimand, such as the target trial emulation framework (Hernán et al., 2022), the causal roadmap (Dang et al.), the estimand framework (Kahan et al., 2024), or the "PICO" (*Population*, *Intervention*, *Control*, and *Outcome*) framework (Riva et al., 2012). As we show below, imprecision on the various components of the causal estimand can change the causal effect or even compromise its validity.

Intervention A well specified intervention is essential for SUTVA (assumption 2.3) to hold (Cole and Frangakis, 2009). In an observational study, we may be tempted to estimate the "effect" of ever *versus* never taking aspirin. This approach is problematic because we do not expect 300 mg of aspirin daily to have the same effect on breast cancer recurrence as 500 mg of aspirin once every year, although both use patterns would be included in the aspirin-ever group. In fact, the intervention "aspirin" is ambiguous without at least the dosage, timing, frequency, and duration of use being specified. The challenge when specifying the interventions is to determine the extent of detail necessary (Hernán, 2016). While dosage, frequency, and duration are commonly acknowledged as factors which may influence the potential outcomes, other details –such as the medication excipients, or whether it is taken during or outside meals– might also potentially be of importance. In theory, interventions should be specified until no different versions may result in different potential outcomes. Determining when this point is met is non-trivial, and requires expert knowledge. The task of unambiguously defining interventions may seem even more complex when dealing with individual attributes such as gender, ethnicity, or body mass index (BMI) (Hernán and Taubman, 2008). Importantly, causal inference is not aimed at identifying causes *per se*, but at estimating the effects of interventions: this is the "no causation without manipulation" paradigm (Holland, 1986). Our causal framework (the potential outcomes framework) does not define, even less measure, the "effect" of a high BMI on an outcome. Rather, this framework allows us to formulate and quantify the effect of specific interventions related to weight loss, such as initiating a hypo-caloric diet program.

Control The choice of the control intervention used as a reference should also be carefully chosen and clearly stated (Rosenbaum, 1999; Malay and Chung, 2012). For example, aspirin may decrease the risk of breast cancer recurrence compared to no therapy but increase the risk of breast cancer recurrence compared to chemotherapy.

Population A causal effect estimate is specific to the population in which it was estimated. Individualized treatment effects as obtained with CATEs can sometimes be applied to a population different from study population, however such generalization can be undermined by different interference structures or treatment variants (Hernán and VanderWeele, 2011). For example, an estimator of the effect of aspirin on breast cancer recurrence derived from a source population may not be directly applicable to an individual

in a target population if there is a limited supply of aspirin pills, resulting in inter-unit interference in the target population, or if the target population uses drugs from a different pharmaceutical company, with different excipients.

Outcome A good specification of the outcome to be evaluation is key to framing a causal question. A first aspect is to enable apple-to-apple comparison across studies. Imagine that aspirin only delays cancer recurrence: we might find a beneficial effect of aspirin if the outcome is recurrence at five years, but no effect if the outcome is recurrence at twenty years. However, another fundamental aspect is the validity of the outcome from a health standpoint. Indeed, considering only recurrence leaves out patients who have died of another cause –a patient who dies can no longer recur. We say that death is a competing event for cancer recurrence (Young et al., 2020). Suppose that aspirin has no specific effect on cancer recurrence but increases the risk of hemorrhagic death. Without accounting for the competing event, aspirin might appear beneficial for cancer recurrence because aspirin-induced hemorrhagic death precludes recurrence. Dedicated statistical methods can handle competing events (Young et al., 2020; Stensrud et al., 2022), in which case these must be explicitly stated for the causal estimand to be unambiguous. An alternative but less precise solution is to consider a composite outcome (Wolbers et al., 2014).

Contrast The ATE, CATE, and all the formulas above are risk differences, based on $Y(1) - Y(0)$. However, there are many other possible *contrasts*, such as risk ratio ($\frac{Y(1)}{Y(0)}$), odds ratio (for binary outcomes), hazard ratio, (Colnet et al., 2023)... An effect might appear highly beneficial on a multiplicative scale but negligible on an additive scale if the baseline risk is low (Forrow et al., 1992; Dj et al., 1993); thus statistical guidelines recommend reporting results on both absolute (*e.g.*, risk difference) and relative (*e.g.*, risk ratio or odds ratio) scales (Schulz et al., 2010; Cuschieri, 2019). In the medical literature, odds ratios are frequent for binary outcomes, and hazard ratios for time-to-event outcomes (Holmberg and Andersen, 2020). However, both of these measures are problematic for causal inference because they are not collapsible : the subgroup causal effects (CATEs) cannot be aggregated into a population causal effect (ATE) (Didelez and Stensrud, 2022; Colnet et al., 2023). In addition, causal interpretation of hazard ratios is challenging because they have a built-in selection bias (Hernán, 2010; Stensrud and Hernán, 2020).

Narrow definitions of the causal estimand –a well-defined population, a homogeneous intervention– help ensuring a valid causal estimation; this is known as *internal validity*. However, it comes at the expense of the *external validity*: the estimated effect is applicable to less situation (Pearce and Vandenbroucke, 2023). Personalized decisions require modeling heterogeneous settings and including important factors of variability as covariates in the CATE.

4.2 Beyond simple confounding: biases can arise from the study design

Even with a precise definition of the target causal estimand and unlimited data, poor study design or data artifacts can introduce systematic biases in the estimation that persist despite the absence of unmeasured confounding or model misspecification (Acton et al., 2023). These biases can take many forms (Spencer et al., 2023), and their detection requires a thorough understanding of the data collection processes and expert knowledge of the underlying causal structure. In the following, we detail a few examples: time alignment failure, measurement bias, and informative losses to follow-up. But many other sources of bias exist (Spencer et al., 2023; Jager et al., 2020; Berrington de González et al., 2024).

Time alignment failures In clinical trials, the start of follow-up (or baseline time), the time of eligibility assessment, and the time of treatment assignment are synchronized. In observational studies, however, these three time points are not naturally defined and should be specified while designing the study. Failure to properly align baseline, eligibility, and intervention assignment times can lead to time alignment failure biases (Hernán et al., 2016).

Consider an observational study of aspirin *versus* control. Suppose the start of follow-up is set at the time of the first breast cancer diagnosis and coincides with the assessment of eligibility criteria. Without random assignment, one must assign patients to the treatment and control group based on their observed patterns of aspirin use. A natural approach would be to assign patients to the treatment arm if they had taken aspirin before breast cancer diagnosis. In this case, treatment allocation occurs before baseline and eligibility screening. This temporal misalignment can introduce *prevalent-user bias* (Danaei et al., 2012). Indeed, if aspirin is effective in preventing breast cancer progression, former aspirin users may

be less likely to develop breast cancer initially and to be included in the analysis. This might bias the study population: patients who are diagnosed with breast cancer despite taking aspirin may represent non-responder individuals. In contrast, individuals in the non-aspirin group would be a mix of resistant and non-resistant patients. Overall, this would lead to underestimating the true effect of aspirin. An alternative approach could be to define the treatment group as patients who start taking aspirin after baseline. In this case, treatment allocation occurs after baseline and eligibility screening. This temporal misalignment can introduce *immortal-time bias* (Suissa, 2008): to be assigned to the aspirin group, patients must survive long enough to start treatment. During this "immortal" period before aspirin is started, outcome events are attributed solely to the control group, which can falsely suggests that aspirin is more beneficial than it actually is. Other variations of time alignment failures exist (Hernán et al., 2016).

As accounting for time is crucial, the "PICO" framework can be extended to "PICOT", where T stands for "time", specifying the duration of the intervention and other temporal patterns of the study (Riva et al., 2012). To avoid defining treatment allocation using post-baseline information, and pitfalls such as immortal-time bias, one solution is the *cloning, censoring, and weighting* approach, within the target trial emulation framework (Hernán et al., 2016). Briefly, it consists of (i) cloning each patient at baseline and assigning one clone to each of the considered interventions, (ii) censoring the clones the first time the patient's behavior is no longer consistent with the assigned intervention, and (iii) weighting the clones to account for the selection bias due to informative censoring induced in the latter step (Maringe et al., 2020; Matthews et al., 2022; Gaber et al., 2024; Huitfeldt et al., 2015). The estimation approaches discussed in section 3 must then use survival models (Ishwaran et al., 2008; Van Belle et al., 2011; Wiegrebe et al., 2024; Alberge et al., 2024).

Measurement bias Measurement bias refers to any bias that results from the process of collecting and preparing the study variables (Hernán and Robins, 2020). We focus on bias arising from mismeasurement of the treatment or outcome, but in principle, bias can also arise from mismeasurement of other variables, including confounding variables. Measurement bias encompasses various pitfalls in data collection, which have been given multiple names, including information bias, recall bias, reverse causation bias, protopathic bias, Berkson bias, interviewer bias, observer bias, and others (Young et al., 2018; Berrington de González et al., 2024; Jager et al., 2020).

Within electronic health records or health claims, a medical condition is generally considered as non-existing in the absence of related care. This could lead to false negative cases and induce measurement bias (Lanes et al., 2015). For example, in healthcare prescription claims, a patient who receives over-the-counter aspirin will be considered an aspirin non-user, and a patient who refuses treatment for cancer recurrence will be considered to have no recurrence. Measurement error is particularly problematic when the measurement error of the treatment (or outcome) is related to the true value of the outcome (or the treatment), in which case it is said to be differential. For example, in retrospective data collected through questionnaires, patients experiencing breast cancer recurrence may be more likely to remember and report aspirin use because they believe it may be related to their condition: this is an example of differential measurement error for the intervention, often referred to as recall bias (Prince, 2012). Another common form of differential measurement error for the intervention, known as reverse causation or protopathic bias (Ri and Ar, 1980; Faillie, 2015), occurs when the intervention is given in response to the first symptoms of the outcome before it is diagnosed or recorded in the dataset. For example, the onset of symptoms of metastatic cancer recurrence could lead patients to take aspirin for bone pain relief before the cancer recurrence is detected. Alternatively, differential measurement error for the outcome can occur if aspirin users tend to visit their primary care physician more often, increasing the likelihood that the cancer recurrence is detected.

The use of proxies for the intervention can also undermine the plausibility of the SUTVA assumption even when the interventions of interest are well-defined (Hernán, 2016). For example, in pharmacy claims, one can only record that the patient received a box of aspirin, not that the patient actually ingested the pills. In this sense, the intervention measured in the dataset is ambiguous: it includes taking any number of pills up to the number of pills in the box. Here, a clinical study would do measure *intention-to-treat effects* (effects of being assigned to medication arm versus the other, whether or the patient complies or not) (Ranganathan et al., 2016). But claims data typical measures buying the treatment, rather than prescription, and thus captures something close to *per-protocol effects* (effects of taking the intervention being assigned).

Measurement bias is both difficult to detect and difficult to manage. Methods to either identify or mitigate measurement bias include, but are not limited to, improving the data collection process, discussing

the direction and magnitude of bias from expert knowledge and prior publications, parametrizing a measurement error model, or performing sensitivity analyses by varying the study design (*e.g.,* using lag periods to reduce the impact of protopathic bias) (Young et al., 2018; Arfè and Corrao, 2016; Schennach, 2016).

Loss to follow-up Loss to follow-up occurs when participants in a study drop out or become unavailable for further data collection before the study is completed, resulting in missing data for the outcome. This type of case, where the outcome is the time to occurrence of a specific event, when observed, can be handled by survival analysis approaches. When informative, loss to follow-up can introduce selection bias, even in randomized trials (Akl et al., 2012; Howe et al., 2016). Selection bias due to informative loss to follow-up occurs when the censoring status variable acts as a collider on the pathway between the intervention and the outcome, or as the descendant of such a collider, thereby introducing collider bias (Hernán and Robins, 2020). For example, consider a randomized controlled trial of aspirin *versus* placebo. Patients in the aspirin arm may experience side effects, leading them to withdraw from the trial at a higher rate than those in the placebo arm. Hence, patients in the aspirin group will become healthier over time than those in the placebo group, and the two arms will no longer be exchangeable, undermining the benefit of the initial randomization.

In most cases, we are interested in estimating the causal effect if no one in the study population were lost to follow-up; that is, the joint effect of aspirin *versus* placebo, under a second intervention that would prevent censoring in both groups. Identification of this causal estimand involves adjustment for informative censoring, typically with weighting methods. This requires new specific causal assumptions to hold (Young et al., 2020).

4.3 Sensitivity analysis to unmeasured confounding

You can't conclude on what you haven't measured. A challenge of causal analysis is that it entirely relies on strong untestable assumptions: conditional positivity and ignorability. This latter assumption states that there must be *no unobserved confounders*, which relies on prior knowledge of the problem. An unobserved covariate can occur in two cases: the confounding variable is not observed or only through a proxy, or the confounding factor is unknown. This latter case is precisely what makes conditional ignorability an untestable assumption.

When some likely confounders an unaccounted for, one approach, known as *sensitivity analysis*, is to assess the properties those missing confounders should have to invalidate the proposed causal result (Robins et al., 2000; Frank et al., 2023). A successful and famous example is the demonstration by Cornfield (Cornfield et al., 1959) that the strength of an unknown genetic factor required to cancel the strong increase in the risk of developing lung cancer when smoking was unrealistic. There are numerous sensitivity-analysis methods, with a variety of underlying additional untestable assumptions, such as the parametric form characterizing the impact of the unobserved confounder or non-parametric bounds simply characterizing the strength of association of the confounder to the treatment and potential outcomes. Recent advances in the field replace those assumptions with non-parametric machine-learning-based approaches (Veitch and Zaveri, 2020; Scharfstein et al., 2021). We refer the reader to more comprehensive reviews for more details (Díaz et al., 2023; Liu et al., 2013; Richardson et al., 2014; Brand et al., 2023). A general conclusion from sensitivity analysis, particularly highlighted by non-parametric bounds such as the widely-used E-value (VanderWeele and Ding, 2017), is that a larger effect is more difficult to invalidate. This is essentially a mathematical formulation of the first criterion, 'strength', on Bradford-Hill's famous criteria for causation (Hill, 1965).

5 Conclusion: machine learning complements but does not replace RCTs

Machine learning can model complex data, such as images, or text. In a causal framework it opens the door to exploiting new data source for evidence-based medical decision-making. Yet, as we have seen, ensuring causal validity from observational data is challenging. The multiple steps, summarized Figure 4, all hide pitfalls. Study design and choice of covariates require a good mastery of the data and associated modeling hypotheses. The choice of machine-learning method also matters (Bouvier et al., 2024; Doutreligne et al., 2023) and may not be obvious: The best predictive model is not necessarily

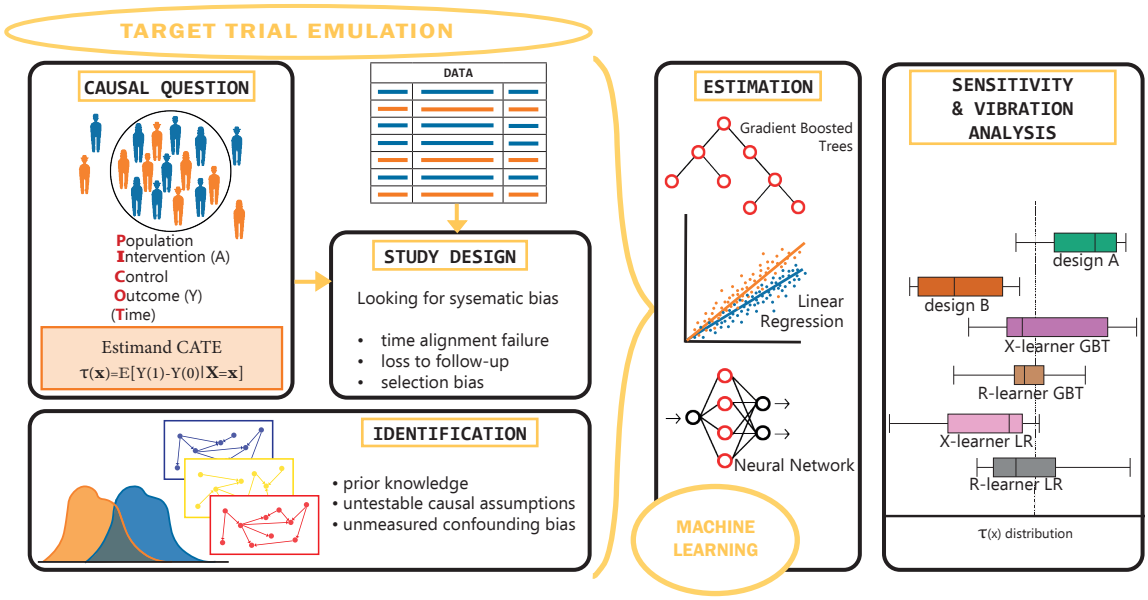

Figure 4: The causal machine learning pipeline.

the causal model, and the best causal model is not necessarily the best predictive one (Doutreligne and Varoquaux, 2023). No machine-learning method can "solve" causal validity with uncontrolled input data, from which they could for instance learning non-causal "shortcuts" (Geirhos et al., 2020). Beyond confounding bias or well-specified estimators, study design, framing well a causal question, is crucial. For these reason, RCTs are typically considered as more solid evidence for causal effects than observational studies (Murad et al., 2016). It can be useful to validate machine-learning causal effects obtained from an observational study by extracting an average effect to compare to an existing RCT (Doutreligne et al., 2023). And yet, for routine decision-making, RCTs are also imperfect (Rothwell, 2005; Deaton and Cartwright, 2018). Their target causal estimand may differ from that of interest: different populations and different interventions erode *external validity*. Individualizing decisions requires estimating a detailed *conditional* effect, which requires larger sample sizes than the typical RCT and can benefit from observing a wide diversity of settings. Finally, decision-making must build on the information available in routine practice, which often differs from that in a clinical study.

There is no magic bullet to building individualized decisions from data; it requires crossing information from RCTs, from observational data at hand, with expert knowledge and machine-learning models.

Acknowledgments We thank Stefan Wager for insightful discussions, and Sarah Abécassis for graphics work on figure 4. JA, JA, and GV acknowledge funding for the project INTERCEPT-T2D by the European Union under the Horizon Europe Programme (Grant Agreement No 101095433), and the PEPR SN SMATCH France 2030 ANR-22-PESN-0003

References

E. K. Acton, A. W. Willis, and S. Hennessy. Core concepts in pharmacoepidemiology: Key biases arising in pharmacoepidemiologic studies. *Pharmacoepidemiology and drug safety*, 32(1):9–18, Jan. 2023. ISSN 1053-8569. . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10204604/.

E. A. Akl, M. Briel, J. J. You, X. Sun, B. C. Johnston, J. W. Busse, S. Mulla, F. Lamontagne, D. Bassler, C. Vera, M. Alshurafa, C. M. Katsios, Q. Zhou, T. Cukierman-Yaffe, A. Gangji, E. J. Mills, S. D. Walter, D. J. Cook, H. J. Schünemann, D. G. Altman, and G. H. Guyatt. Potential impact on

estimated treatment effects of information lost to follow-up in randomised controlled trials (LOST-IT): systematic review. *BMJ*, 344:e2809, May 2012. ISSN 1756-1833. . URL https://www.bmj.com/content/344/bmj.e2809. Publisher: British Medical Journal Publishing Group Section: Research.

J. Alberge, V. Maladière, O. Grisel, J. Abécassis, and G. Varoquaux. Survival models: Proper scoring rule and stochastic optimization with competing risks. *arXiv preprint arXiv:2410.16765*, 2024.

A. Arfè and G. Corrao. The lag-time approach improved drug-outcome association estimates in presence of protopathic bias. *Journal of Clinical Epidemiology*, 78:101–107, Oct. 2016. ISSN 1878-5921. .

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

J. Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946.

A. Berrington de González, D. B. Richardson, and M. K. Schubauer-Berigan, editors. *Statistical methods in cancer research, Volume V. Bias assessment in case–control and cohort studies for hazard identification*, volume 171 of *IARC Scientific Publications*. International Agency for Research on Cancer, Lyon, France, 2024. URL https://publications.iarc.who.int/634.

T. Blakely, J. Lynch, K. Simons, R. Bentley, and S. Rose. Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference. *International journal of epidemiology*, 49 (6):2058–2064, 2020.

F. Bouvier, E. Peyrot, A. Balendran, C. Ségalas, I. Roberts, F. Petit, and R. Porcher. Do machine learning methods lead to similar individualized treatment rules? a comparison study on real data. *Statistics in Medicine*, 43(11):2043–2061, 2024.

J. E. Brand, X. Zhou, and Y. Xie. Recent developments in causal inference and machine learning. *Annual Review of Sociology*, 49(1):81–110, 2023.

L. Breiman. Stacked regressions. *Machine learning*, 24:49–64, 1996.

C. R. Charig, D. R. Webb, S. R. Payne, and J. E. Wickham. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *Br Med J (Clin Res Ed)*, 292(6524):879–882, 1986.

W. Y. Chen and M. D. Holmes. Role of Aspirin in Breast Cancer Survival. *Current Oncology Reports*, 19(7):48, June 2017. ISSN 1534-6269. . URL https://doi.org/10.1007/s11912-017-0605-6.

W. Y. Chen, K. V. Ballman, A. H. Partridge, O. M. Hahn, F. M. Briccetti, W. J. Irvin, B. Symington, K. Visvanathan, P. R. Pohlmann, T. H. Openshaw, A. Weiss, E. P. Winer, L. A. Carey, and M. D. Holmes. Aspirin vs Placebo as Adjuvant Therapy for Breast Cancer: The Alliance A011502 Randomized Trial. *JAMA*, 331(20):1714–1721, May 2024. ISSN 0098-7484. . URL https://doi.org/10.1001/jama.2024.4840.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

C. Cinelli, A. Forney, and J. Pearl. A crash course in good and bad controls. *Sociological Methods & Research*, 53(3):1071–1104, 2024.

S. R. Cole and C. E. Frangakis. The Consistency Statement in Causal Inference: A Definition or an Assumption? *Epidemiology*, 20(1):3, Jan. 2009. ISSN 1044-3983. . URL https://journals.lww.com/epidem/fulltext/2009/01000/the_consistency_statement_in_causal_inference__a.3.aspx.

B. Colnet, J. Josse, G. Varoquaux, and E. Scornet. Risk ratio, odds ratio, risk difference... which causal measure is easier to generalize? *arXiv preprint arXiv:2303.16008*, 2023.

J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22(1):173–203, 1959.

W. H. Crown. Real-world evidence, causal inference, and machine learning. *Value in Health*, 22(5): 587–592, 2019.

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

A. Curth and M. Van der Schaar. On inductive biases for heterogeneous treatment effect estimation. *Advances in Neural Information Processing Systems*, 34:15883–15894, 2021.

A. Curth, R. W. Peck, E. McKinney, J. Weatherall, and M. van Der Schaar. Using machine learning to individualize treatment effect estimation: Challenges and opportunities. *Clinical Pharmacology & Therapeutics*, 115(4):710–719, 2024.

S. Cuschieri. The STROBE guidelines. *Saudi Journal of Anaesthesia*, 13(Suppl 1):S31–S34, Apr. 2019. ISSN 1658-354X. . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6398292/.

G. Danaei, M. Tavakkoli, and M. A. Hernán. Bias in observational studies of prevalent users: lessons for comparative effectiveness research from a meta-analysis of statins. *American Journal of Epidemiology*, 175(4):250–262, Feb. 2012. ISSN 1476-6256. .

L. E. Dang, S. Gruber, H. Lee, I. J. Dahabreh, E. A. Stuart, B. D. Williamson, R. Wyss, I. Díaz, D. Ghosh, E. Kıcıman, D. Alemayehu, K. L. Hoffman, C. Y. Vossen, R. A. Huml, H. Ravn, K. Kvist, R. Pratley, M.-C. Shih, G. Pennello, D. Martin, S. P. Waddy, C. E. Barr, M. Akacha, J. B. Buse, M. van der Laan, and M. Petersen. A causal roadmap for generating high-quality real-world evidence. *Journal of Clinical and Translational Science*, 7(1):e212. ISSN 2059-8661. . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10603361/.

A. Deaton and N. Cartwright. Understanding and misunderstanding randomized controlled trials. *Social science & medicine*, 210:2–21, 2018.

I. Díaz, H. Lee, E. Kıcıman, E. J. Schenck, M. Akacha, D. Follman, and D. Ghosh. Sensitivity analysis for causality in observational studies for regulatory science. *Journal of Clinical and Translational Science*, 7(1):e267, 2023.

V. Didelez and M. J. Stensrud. On the logic of collapsibility for causal effect measures. *Biometrical Journal. Biometrische Zeitschrift*, 64(2):235–242, Feb. 2022. ISSN 1521-4036. .

M. Dj, B. Ja, J. S, W. Jw, and R. Jm. The framing effect of relative and absolute risk. *Journal of general internal medicine*, 8(10), Oct. 1993. ISSN 0884-8734. . URL https://pubmed.ncbi.nlm.nih.gov/8271086/. Publisher: J Gen Intern Med.

J. Dockès, G. Varoquaux, and J.-B. Poline. Preventing dataset shift from breaking machine-learning biomarkers. *GigaScience*, 10(9):giab055, 2021.

M. Doutreligne and G. Varoquaux. How to select predictive models for causal inference? *arXiv preprint arXiv:2302.00370*, 2023.

M. Doutreligne, T. Struja, J. Abecassis, C. Morgand, G. Varoquaux, and L. A. Celi. Step-by-step causal analysis of electronic health records to ground decision making. 2023.

A. D'Amour, P. Ding, A. Feller, L. Lei, and J. Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

J.-L. Faillie. Indication bias or protopathic bias? *British Journal of Clinical Pharmacology*, 80(4):779–780, Oct. 2015. ISSN 0306-5251. . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4594717/.

M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, and F. Hutter. Auto-sklearn 2.0: Hands-free automl via meta-learning. *Journal of Machine Learning Research*, 23(261):1–61, 2022.

L. Forrow, W. C. Taylor, and R. M. Arnold. Absolutely relative: how research results are summarized can affect treatment decisions. *The American Journal of Medicine*, 92(2):121–124, Feb. 1992. ISSN 0002-9343. .

D. J. Foster and V. Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908, 2023.

K. A. Frank, Q. Lin, R. Xu, S. Maroulis, and A. Mueller. Quantifying the robustness of causal inferences: Sensitivity analysis for pragmatic social science. *Social Science Research*, 110:102815, 2023.

C. E. Gaber, K. A. Hanson, S. Kim, J. L. Lund, T. A. Lee, and E. J. Murray. The Clone-Censor-Weight Method in Pharmacoepidemiologic Research: Foundations and Methodological Implementation. *Current Epidemiology Reports*, 11(3):164–174, Sept. 2024. ISSN 2196-2995. . URL https://doi.org/10.1007/s40471-024-00346-2.

R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

S. Greenland and B. Brumback. An overview of relations among causal modelling methods. *International journal of epidemiology*, 31(5):1030–1037, 2002.

G. J. Griffith, T. T. Morris, M. J. Tudball, A. Herbert, G. Mancano, L. Pike, G. C. Sharp, J. Sterne, T. M. Palmer, G. Davey Smith, et al. Collider bias undermines our understanding of covid-19 disease risk and severity. *Nature communications*, 11(1):5749, 2020.

L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

P. R. Hahn, J. S. Murray, and C. M. Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3): 965–1056, 2020.

M. A. Hernán and J. M. Robins. Causal inference: What if, 2020.

M. A. Hernán. The Hazards of Hazard Ratios. *Epidemiology (Cambridge, Mass.)*, 21(1):13–15, Jan. 2010. ISSN 1044-3983. . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3653612/.

M. A. Hernán. Does water kill? A call for less casual causal inferences. *Annals of Epidemiology*, 26(10): 674–680, Oct. 2016. ISSN 1047-2797. . URL https://www.sciencedirect.com/science/article/pii/S1047279716302800.

M. A. Hernán and J. M. Robins. *Causal Inference: What If*. Boca raton: Chapman & hall/crc. edition, 2020.

M. A. Hernán and S. L. Taubman. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal of Obesity*, 32(3):S8–S14, Aug. 2008. ISSN 1476-5497. . URL https://www.nature.com/articles/ijo200882. Publisher: Nature Publishing Group.

M. A. Hernán and T. J. VanderWeele. Compound Treatments and Transportability of Causal Inference. *Epidemiology*, 22(3):368, May 2011. ISSN 1044-3983. . URL https://journals.lww.com/epidem/fulltext/2011/05000/compound_treatments_and_transportability_of_causal.18.aspx.

M. A. Hernán, B. C. Sauer, S. Hernández-Díaz, R. Platt, and I. Shrier. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *Journal of Clinical Epidemiology*, 79:70–75, Nov. 2016. ISSN 1878-5921. .

M. A. Hernán, J. Hsu, and B. Healy. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *CHANCE*, 32(1):42–49, Jan. 2019. ISSN 0933-2480. . URL https://doi.org/10.1080/09332480.2019.1579578. Publisher: ASA Website _eprint: https://doi.org/10.1080/09332480.2019.1579578.

M. A. Hernán, W. Wang, and D. E. Leaf. Target Trial Emulation: A Framework for Causal Inference From Observational Data. *JAMA*, 328(24):2446–2447, Dec. 2022. ISSN 1538-3598. .

A. B. Hill. The environment and disease: association or causation?, 1965.

S. R. Hoffman, N. Gangan, X. Chen, J. L. Smith, A. Tave, Y. Yang, C. L. Crowe, S. dosReis, and M. Grabner. A step-by-step guide to causal study design using real-world data. *Health Services and Outcomes Research Methodology*, June 2024. ISSN 1572-9400. . URL https://doi.org/10.1007/s10742-024-00333-6.

P. W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396): 945–960, 1986.

M. J. Holmberg and L. W. Andersen. Estimating Risk Ratios and Risk Differences: Alternatives to Odds Ratios. *JAMA*, 324(11):1098–1099, Sept. 2020. ISSN 0098-7484. . URL https://doi.org/10.1001/jama.2020.12698.

C. J. Howe, S. R. Cole, B. Lau, S. Napravnik, and J. J. Eron. Selection bias due to loss to follow up in cohort studies. *Epidemiology (Cambridge, Mass.)*, 27(1):91–97, Jan. 2016. ISSN 1044-3983. . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5008911/.

M. Huber. An introduction to causal discovery. *Swiss Journal of Economics and Statistics*, 160(1):14, Oct. 2024. ISSN 2235-6282. . URL https://doi.org/10.1186/s41937-024-00131-4.

A. Huitfeldt, M. Kalager, J. M. Robins, G. Hoff, and M. A. Hernán. Methods to Estimate the Comparative Effectiveness of Clinical Strategies that Administer the Same Intervention at Different Times. *Current epidemiology reports*, 2(3):149–161, Sept. 2015. ISSN 2196-2995. . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4646164/.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. 2008.

K. J. Jager, G. Tripepi, N. C. Chesnaye, F. W. Dekker, C. Zoccali, and V. S. Stel. Where to look for the most frequent biases? *Nephrology (Carlton, Vic.)*, 25(6):435–441, June 2020. ISSN 1320-5358. . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7318122/.

S. A. Julious and M. A. Mullee. Confounding and simpson's paradox. *Bmj*, 309(6967):1480–1481, 1994.

B. C. Kahan, J. Hindley, M. Edwards, S. Cro, and T. P. Morris. The estimands framework: a primer on the ICH E9(R1) addendum. *BMJ*, 384:e076316, Jan. 2024. ISSN 1756-1833. . URL https://www.bmj.

com/content/384/bmj-2023-076316. Publisher: British Medical Journal Publishing Group Section: Research Methods & Reporting.

E. H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

E. H. Kennedy, S. Balakrishnan, J. M. Robins, and L. Wasserman. Minimax rates for heterogeneous causal effect estimation. *The Annals of Statistics*, 52(2):793–816, 2024.

S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

S. Lanes, J. S. Brown, K. Haynes, M. F. Pollack, and A. M. Walker. Identifying health outcomes in healthcare databases. *Pharmacoepidemiology and Drug Safety*, 24(10):1009–1016, 2015. ISSN 1099-1557. . URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pds.3856. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pds.3856.

F. Li, L. E. Thomas, and F. Li. Addressing extreme propensity scores via the overlap weights. *American journal of epidemiology*, 188(1):250–257, 2019.

F. Liu. Data science methods for real-world evidence generation in real-world data. *Annual Review of Biomedical Data Science*, 7, 2024.

W. Liu, S. J. Kuramoto, and E. A. Stuart. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention science*, 14:570–580, 2013.

C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

D. P. MacKinnon and S. J. Lamp. A unification of mediator, confounder, and collider effects. *Prevention Science*, 22(8):1185–1193, 2021.

S. Malay and K. C. Chung. The Choice of Controls for Providing Validity and Evidence in Clinical Research. *Plastic and reconstructive surgery*, 130(4):959–965, Oct. 2012. ISSN 0032-1052. . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3461178/.

C. Maringe, S. Benitez Majano, A. Exarchakou, M. Smith, B. Rachet, A. Belot, and C. Leyrat. Reflection on modern methods: trial emulation in the presence of immortal-time bias. Assessing the benefit of major surgery for elderly lung cancer patients using observational data. *International Journal of Epidemiology*, 49(5):1719–1729, Oct. 2020. ISSN 0300-5771. . URL https://doi.org/10.1093/ije/dyaa057.

A. A. Matthews, G. Danaei, N. Islam, and T. Kurth. Target trial emulation: applying principles of randomised trials to observational studies. *BMJ*, 378:e071108, Aug. 2022. ISSN 1756-1833. . URL https://www.bmj.com/content/378/bmj-2022-071108. Publisher: British Medical Journal Publishing Group Section: Research.

M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, and M. Lucic. Revisiting the calibration of modern neural networks. *NeurIPS*, 34, 2021.

M. H. Murad, N. Asi, M. Alsawas, and F. Alahdab. New evidence pyramid. *BMJ Evidence-Based Medicine*, 21(4):125–127, 2016.

A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2): 299–319, 2021.

M. Oberst, F. Johansson, D. Wei, T. Gao, G. Brat, D. Sontag, and K. Varshney. Characterization of overlap in observational studies. In *International Conference on Artificial Intelligence and Statistics*, pages 788–798. PMLR, 2020.

C. J. Patel, B. Burford, and J. P. Ioannidis. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of clinical epidemiology*, 68(9): 1046–1058, 2015.

N. Pearce and J. P. Vandenbroucke. Are target trial emulations the gold standard for observational studies? *Epidemiology*, 34(5):614–618, 2023.

A. Perez-Lebel, M. Le Morvan, and G. Varoquaux. Beyond calibration: estimating the grouping loss of modern neural networks. *ICLR*, 2022.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

M. Prince. 9 - Epidemiology. In P. Wright, J. Stern, and M. Phelan, editors, *Core Psychiatry (Third Edition)*, pages 115–129. W.B. Saunders, Oxford, Jan. 2012. ISBN 978-0-7020-3397-1. . URL https://www.sciencedirect.com/science/article/pii/B9780702033971000094.

M. Prosperi, Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He, S. Rich, M. Wang, I. E. Buchan, and J. Bian. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, 2(7):369–375, 2020.

P. Ranganathan, C. Pramesh, and R. Aggarwal. Common pitfalls in statistical analysis: Intention-to-treat versus per-protocol analysis. *Perspectives in clinical research*, 7(3):144–146, 2016.

H. Ri and F. Ar. The problem of "protopathic bias" in case-control studies. *The American journal of medicine*, 68(2), Feb. 1980. ISSN 0002-9343. . URL https://pubmed.ncbi.nlm.nih.gov/7355896/. Publisher: Am J Med.

A. Richardson, M. G. Hudgens, P. B. Gilbert, and J. P. Fine. Nonparametric bounds and sensitivity analysis of treatment effects. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(4):596, 2014.

J. J. Riva, K. M. Malik, S. J. Burnie, A. R. Endicott, and J. W. Busse. What is your research question? An introduction to the PICOT format for clinicians. *The Journal of the Canadian Chiropractic Association*, 56(3):167–171, Sept. 2012. ISSN 0008-3194. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3430448/.

J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

J. M. Robins, A. Rotnitzky, and D. O. Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 1–94. Springer, 2000.

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, (4):931–954, 1988.

P. R. Rosenbaum. Choice as an Alternative to Control in Observational Studies. *Statistical Science*, 14(3):259–278, 1999. ISSN 0883-4237. URL https://www.jstor.org/stable/2676761. Publisher: Institute of Mathematical Statistics.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

P. M. Rothwell. External validity of randomised controlled trials:"to whom do the results of this trial apply?". *The Lancet*, 365(9453):82–93, 2005.

D. O. Scharfstein, R. Nabi, E. H. Kennedy, M.-Y. Huang, M. Bonvini, and M. Smid. Semiparametric sensitivity analysis: Unmeasured confounding in observational studies. *arXiv preprint arXiv:2104.08300*, 2021.

S. M. Schennach. Recent Advances in the Measurement Error Literature. *Annual Review of Economics*, 8(Volume 8, 2016):341–377, Oct. 2016. ISSN 1941-1383, 1941-1391. . URL https://www.annualreviews.org/content/journals/10.1146/annurev-economics-080315-015058. Publisher: Annual Reviews.

E. F. Schisterman, S. R. Cole, and R. W. Platt. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*, 20(4):488–495, 2009.

K. F. Schulz, D. G. Altman, D. Moher, and the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMC Medicine*, 8(1):18, Mar. 2010. ISSN 1741-7015. . URL https://doi.org/10.1186/1741-7015-8-18.

E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

E. A. Spencer, C. Heneghan, and J. K. Aronson. Catalog of bias, 2023. URL https://www.catalogueofbiases.org. Catalogue of Bias Collaboration.

M. J. Stensrud and M. A. Hernán. Why Test for Proportional Hazards? *JAMA*, 323(14):1401–1402, Apr. 2020. ISSN 1538-3598. .

M. J. Stensrud, J. G. Young, V. Didelez, J. M. Robins, and M. A. Hernán. Separable Effects for Causal Inference in the Presence of Competing Events. *Journal of the American Statistical Association*, 117(537):175–183, Jan. 2022. ISSN 0162-1459. . URL https://doi.org/10.1080/01621459.2020.1765783. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2020.1765783.

S. Suissa. Immortal Time Bias in Pharmacoepidemiology. *American Journal of Epidemiology*, 167(4):492–499, Feb. 2008. ISSN 0002-9262. . URL https://doi.org/10.1093/aje/kwm324.

P. W. Tennant, E. J. Murray, K. F. Arnold, L. Berrie, M. P. Fox, S. C. Gadd, W. J. Harrison, C. Keeble, L. R. Ranker, J. Textor, et al. Use of directed acyclic graphs (dags) to identify confounders in applied health research: review and recommendations. *International journal of epidemiology*, 50(2):620–632, 2021.

J. Textor, B. Van der Zander, M. S. Gilthorpe, M. Liśkiewicz, and G. T. Ellison. Robust causal inference using directed acyclic graphs: the r package 'dagitty'. *International journal of epidemiology*, 45(6):1887–1894, 2016.

V. Van Belle, K. Pelckmans, S. Van Huffel, and J. A. Suykens. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial intelligence in medicine*, 53(2):107–118, 2011.

M. J. van der Laan and A. R. Luedtke. Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome. 2014.

M. J. Van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

T. J. VanderWeele and P. Ding. Sensitivity analysis in observational research: introducing the e-value. *Annals of internal medicine*, 167(4):268–274, 2017.

G. Varoquaux and O. Colliot. Evaluating machine learning models and their diagnostic value. *Machine learning for brain disorders*, pages 601–630, 2023.

V. Veitch and A. Zaveri. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. *Advances in neural information processing systems*, 33:10999–11009, 2020.

V. Veitch, D. Sridhar, and D. Blei. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pages 919–928. PMLR, 2020.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

S. Wiegrebe, P. Kopper, R. Sonabend, B. Bischl, and A. Bender. Deep learning for survival analysis: a review. *Artificial Intelligence Review*, 57(3):65, 2024.

M. Wolbers, M. T. Koller, V. S. Stel, B. Schaer, K. J. Jager, K. Leffondre, and G. Heinze. Competing risks analyses: objectives and approaches. *European heart journal*, 35(42):2936–2941, 2014.

J. C. Young, M. M. Conover, and M. J. Funk. Measurement error and misclassification in electronic medical records: methods to mitigate bias. *Current epidemiology reports*, 5(4):343–356, Dec. 2018. ISSN 2196-2995. . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9141310/.

J. G. Young, M. J. Stensrud, E. J. T. Tchetgen, and M. A. Hernán. A causal framework for classical statistical estimands in failure time settings with competing events. *Statistics in medicine*, 39(8): 1199–1236, Apr. 2020. ISSN 0277-6715. . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7811594/.