TREATMENT EFFECT: A CRITIQUE

H. S. BATTEY AND CHARLOTTE EDGAR

SUMMARY. Two broad positions within statistics define a treatment effect, on the one hand, as a parameter of a statistical model, and on the other, as a contrast in potential or counterfactual outcomes under the different treatment regimes. This short expository paper presents some simple but consequential insights on the two formulations, contrasting the answers under the most favourable fictitious idealisation for the counterfactual framework. These observations clarify the relationship between Fisherian model-based inference and popular counterfactual formulations, and emphasise concerns, raised by Cox (1958b, 2012) and others, regarding the suitability of model-free definitions as targets of inference when scientific conclusions are intended to generalise beyond the observed sample. Parts of the paper are necessarily controversial; we follow Cox (1958a) in not putting these forward in any dogmatic spirit.

Some key words: causation; counterfactuals; foundations of modelling; scientific practice.

1. INTRODUCTION

When the aim is scientific understanding rather than immediate predictive success, establishing causation is usually the idealised objective even when the word is not explicitly used. The terminology *causal inference* has, however, come to be principally associated with a specific formulation based on counterfactuals, in the vein of Rubin (1974) and Rosenbaum and Rubin (1983). This has gained popularity as a basis for methodological development, with many recent contributions emphasising the property of double robustness first introduced by Robins *et al.* (1994). Of two primary aspects that are open to discussion, we focus here on the way in which the effect of potential causes is measured, revisiting the fundamental question *what is a treatment effect?* (McCullagh, 2002).

The model-free definition of treatment effect implicit in the counterfactual formulations differs strikingly from a framing in which the treatment effect is defined as a parameter of a statistical model, in the vein of Fisher (1922), who appears to have been the first to use the word *parameter* in the general modern sense (see Stigler, 1976). In view of this, and also of Bailey's (2017) correction to a historical misnomer (see §4.4), it is appropriate to associate with Fisher's name the model-based definition of treatment effect rather than a definition implicit in what is sometimes referred to as *Fisher's sharp null hypothesis*. The model-based definition encompasses a style of discussion often suggestive of causation and reflected in regression-type formulations (e.g. Cox, 1958b, 1968, 1972, 2007; McCullagh and Nelder, 1989).

We do not claim any great originality for the points made; the contribution is to formulate examples in which the issues are isolated in the most incisive form. Our position echoes concerns raised by Cox (1958b, 2012), who questioned the suitability of model-free averages as targets of inference when conclusions are intended to generalise beyond the observed sample. See also McCullagh (2022, pp.397–399) for a closely related discussion emphasising different aspects.

Department of Mathematics, Imperial College London, 180 Queen's Gate, London SW7 2AZ, UK.
h.battey@imperial.ac.uk and c.edgar24@imperial.ac.uk.

2. Model-based treatment effect

2.1. **Introduction.** Parametric statistical inference starts from a provisional assumption that data correspond to random variables drawn from an unknown probability distribution, and that this distribution belongs to a parametrised family $\mathcal{P}_\Theta := \{P_\theta : \theta \in \Theta\}$, the statistical model. Dependence or independence, perhaps conditional, between observations is reflected in $\mathcal{P}_\Theta$. While there is considerable flexibility in how a statistical model might be constructed, a core principle is that $\mathcal{P}_\Theta$ should stress commonalities between observational or experimental units, the purpose being to provide subject-matter understanding in such a way that any conclusions drawn remain relevant to units different from those already observed. Such commonalities are encapsulated in stable parameters, of which a treatment parameter, quantifying treatment effect, is one example.

McCullagh (2002) formalised some principles underpinning the construction of sensible statistical models, noting that the textbook definition does not incorporate by default aspects usually invoked, with the implication that four "plainly absurd" examples are in principle valid. His work can be viewed as an attempt to put scientific common sense on a more rigorous mathematical footing and stems to some extent from an earlier technical report (McCullagh, 1999) establishing a group theoretic perspective on generalised linear models via exponential tilting. McCullagh (2022, Ch. 11 and 14) gives an accessible exposition of some of the main ideas. An important point is that this formalisation is detached from any causal mechanism; it is rather the context that determines whether causation can be securely established.

2.2. **McCullagh's formalisation.** In the regression formulation, the treatment has an effect, not on the outcome, but on the probability distribution of the outcome. McCullagh (2022, p. 231, p. 399) expressed this as a transformation induced by the action of a group $\mathcal{G}$ on $\{P_\gamma : \gamma \in \Gamma\}$, the set of baseline distributions for the outcome prior to possible treatment. In the simplest models the group acts on the baseline distributions via the baseline parameter space Γ as $\mathcal{G} \times \Gamma \to \Gamma$, i.e. for any $g \in \mathcal{G}$ and $\gamma \in \Gamma$, $(g, \gamma) \to g\gamma \in \Gamma$, where $g\gamma$ is the action of g on γ. As two simple examples: if $\Gamma = \mathbb{R}$, a natural group action is addition, $g = (\psi, +)$ say, so that $g\gamma = \gamma + \psi$; if $\Gamma = \mathbb{R}^+$, the natural group action is multiplication, $g = (\psi, \times)$ say, and $g\gamma = \gamma\psi$. In each case the null treatment effect is the identity element of the group, i.e. addition of zero in the first example and multiplication by one in the second.

In what follows, we will use the imprecise terminology *individual* to refer to an experimental or observational unit of any kind, which may, for instance, be nested within individuals. This terminology reflects the conventional and suggestive indexing by i. The effect of the treatment is to modify the response distribution, altering the control distribution by the same group action for all treated individuals. Implicit in this is the requirement that outcome distributions are notionally compared for individuals i and i', perhaps hypothetical, that are comparable in the sense that for all events A in the sample space for the outcome,

$$\begin{aligned} \mathrm{pr}(Y_i \in A \mid T = 0) &= \mathrm{pr}(Y_{i'} \in A \mid T = 0) \\ \mathrm{pr}(Y_i \in A \mid T = 1) &= \mathrm{pr}(Y_{i'} \in A \mid T = 1), \end{aligned} \tag{2.1}$$

where here and henceforth we restrict attention to a binary treatment variable taking values $t \in \{0, 1\}$. These probabilities in general depend on i: in the simplest parametrised models within this simplest setting, (2.1) implies that i and i' share the same baseline parameter $\gamma_i = \gamma_{i'}$, and if both are treated, both are perturbed from baseline in the same way, i.e. in the first line $P_{\gamma_i} = P_{\gamma_i'}$ and in the second $P_{g\gamma_i} = P_{g\gamma_i'}$. Dependence of the probabilities in (2.1) on intrinsic features, w say, recovers standard regression formulations. Specifically, if $\Gamma = \mathbb{R}$, a popular model would be the linear regression with $\gamma_i = w_i^{\mathrm{T}}\beta$, while if $\Gamma = \mathbb{R}^+$,

the common modelling assumption $\gamma_i = \exp(w_i^{\mathrm{T}}\beta)$ would recover other popular regression formulations. There is no assumption in (2.1) about the mechanism through which the treatment is assigned. While the group element defining the treatment effect is stable over individuals, the control distribution is individual-specific in view of γ_i. Interaction of treatment with intrinsic variables is discussed on pp. 232-240 of McCullagh (2022), allowing some variation of treatment effects by unit while respecting the stability requirement.

3. Model-free treatment effect and counterfactuals

In model-free settings there is no outcome model in which to embed a natural and stable treatment parameter. For a binary treatment, a common choice is an average difference in potential outcomes, the simplest model-free counterfactual treatment effect being

$$\tau_n := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_i(Y_i(1) - Y_i(0)), \tag{3.1}$$

where $Y_i(t)$ is the outcome on individual i for values $t \in \{0, 1\}$ of the treatment indicator. The form (3.1) will be taken as illustrative of the general issues to be presented but a model-free treatment effect need not be defined in terms of a difference on this scale. Ratios (log differences) and other comparisons are possible in the counterfactual framework too.

The terminology *potential outcomes* is favoured by Imbens and Rubin (2015), embracing both the factual and counterfactual version of the outcome prior to observation. Some clarification is required regarding the expectation in (3.1). Imbens and Rubin (2015) consider two frameworks: a finite population setting and a so-called superpopulation setting. In their corresponding definitions of treatment effect, n is the population size in the first formulation and a sample size in the second. Our discussion here relates to the superpopulation setting, otherwise there is no hope for generalising conclusions beyond the individuals who have been observed, to our mind a core objective for statistical work. In both settings Imbens and Rubin (2015) consider the potential outcomes to be deterministic functions of the treatment received, possibly different for each individual, and consider randomness as arising from treatment assignment in the finite sample case, and from random sampling in the superpopulation setting. Hernán and Robins (2020), by contrast, consider the potential outcomes themselves to be random, and this is what we have in mind in (3.1).

Since, for a given i, only one of $Y_i(1)$ or $Y_i(0)$ is observable, two broad approaches are common, based on assumptions about the unobservable "cross world" of the extended counterfactual process: one is to impute the counterfactual on the basis of covariate information, requiring an outcome model; the other is propensity-score reweighting (Rosenbaum and Rubin, 1983), which models the treatment-assignment mechanism instead. Under the strong assumption of no unmeasured confounding and several further assumptions, it can be shown that the resulting observable statistic, $\hat{\tau}_n$ say, estimates τ_n.

Although Imbens and Rubin (2015, Chapter 8) refer to model-based approaches, such models are only used as a means for performing inference on (3.1) or a similarly defined model-free treatment effect. Their usage of this terminology is thus different from that of §2, in the sense that the role of the model is to impute counterfactual outcomes in order to estimate (3.1).

4. Some prompts for reflection

4.1. **Two aspects.** There are two distinct types of consideration regarding these strikingly different definitions of treatment effect. One concerns the security of causal claims and the strength of untestable assumptions, and the other concerns the way in which the effect of the potential cause is measured. The first of these has been covered elsewhere in the literature, notably by Dawid (2000) and McCullagh (2022, p. 244). The criticism is that a formulation in terms of objects that are inherently unobservable can never be falsified, and

therefore violates core principles of the scientific method, as put forward by e.g. Popper (1963). We will not repeat any of this earlier discussion, focussing instead on the second consideration, which can be decoupled from the first under a fictitious idealisation for the counterfactuals.

4.2. **Fictitious idealisation.** To detach concerns over the use of counterfactuals from those over what is being measured, we suppose, notionally, that both potential outcomes are observable and independent, conditional on the information identifying the individual. This fictitious idealisation is the most favourable framework for the counterfactual formulation and permits isolation of key issues in an incisive form.

Under the fictitious idealisation, the counterfactual formulation reduces to a matched-pair problem in which each individual is twinned with a perfect copy of itself. From individual i and his or her fictitious twin, one of the two is randomised to treatment, the other to control, yielding notionally observable outcomes (Y_{i1}, Y_{i0}). Given an outcome model for (Y_{i1}, Y_{i0}) with a stable treatment parameter, the question we seek to address is to what extent the model-free definition of treatment effect (3.1) recovers it, and how this depends on the structure of the outcome model. If the answers differ appreciably under simple idealised settings where a stable treatment effect is unambiguously defined, and where all practical difficulties associated with counterfactuals are notionally absent, this must raise concerns over the relevance of (3.1) in more difficult contexts. We defer to readers the judgement over whether our examples offer unambiguous definitions of treatment effect; we believe that they do.

4.3. **Five examples.** The simplest example in which the two formulations give identical answers is a linear model of the form $Y_{i0} = \gamma_i + \varepsilon_i$ and $Y_{i1} = \gamma_i + \Delta + \varepsilon_i'$ with ε_i and ε_i' independent mean-zero error terms. This encompasses the linear regression model with intrinsic variables w by taking $\gamma_i = w_i^{\mathrm{T}}\beta$. Implicit in the model just presented is an assumption of unit treatment additivity that underpinned R. A. Fisher's early work on the design of experiments. The model-free treatment effect (3.1) recovers the model-based treatment effect Δ in this case. This example is essentially the same as that discussed by Imbens and Rubin (2015, pp. 122).

A similarly simple example takes Y_{i0} and Y_{i1} as exponentially distributed of rates γ_i and $\gamma_i\theta$ respectively, so that the model-based treatment effect is multiplicative on the rate scale. If a difference of the form (3.1) is required, it is appropriate, since the outcomes are positive, to construct the model-free treatment effect (3.1) after logarithmic transformation, giving $\tau_n = -\log\theta$, which is the model-based treatment effect θ converted from the rate scale to the log mean scale. The two approaches are again compatible.

The previous example can be elaborated by extending the outcome distribution to Weibull with shape parameter α, the rate parameters being specified as above, with multiplicative treatment effect θ on the rate scale. Since the density function of the ratio $Z_i = Y_{i1}/Y_{i0}$ is

$$f_Z(z) = \frac{\alpha\theta z^{\alpha-1}}{(1+\theta z^{\alpha})^2}, \quad z \geq 0,$$

the definition (3.1) yields, after logarithmic transformation,

$$\tau_n = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\log Y_{i1} - \log Y_{i0}) = \int_0^{\infty}\frac{\log(z)\alpha\theta z^{\alpha-1}}{(1+\theta z^{\alpha})^2}dz = -\frac{\log\theta}{\alpha},$$

showing that the true treatment effect θ can be recovered only up to the sign of $\log\theta$, unless α is known, as in the previous exponential example with $\alpha = 1$. This is an example in which the answers are semi-compatible, being qualitatively reasonable and stable across samples, but giving different numerical values even when converted to a compatible scale.

We now present two examples in which the disagreement seems more severe. Suppose that Y_{i0} and Y_{i1} are exponentially distributed with rates γ_i and $\gamma_i + \Delta$ respectively. This model is physically natural when the outcome is, say, the time to the first event in a Poisson process defined as the superposition of two separate Poisson processes, the baseline process with rate $\gamma_i > 0$ and the treatment process with rate $\Delta > 0$. In other words, the treatment, or exposure, must necessarily increase or not decrease the rate at which events occur. In principle, one could extend the model to allow $\Delta > -\min_i \gamma_i$ but the sample-dependent parameter space, while unproblematic for conditional inference on Δ (see Battey and Cox, 2020, §4), violates McCullagh's (2002) stability axiom and its group-theoretic formalism, and is thus less convincing for isolating the aberration of interest.

Since the density function of the ratio $Z_i = Y_{i1}/Y_{i0}$ is

$$f_{Z_i}(z) = \frac{\gamma_i(\gamma_i + \Delta)}{(\gamma_i + (\gamma_i + \Delta)z)^2}, \quad z \geq 0,$$

equation (3.1) yields, after logarithmic transformation

$$\tau_n = \frac{1}{n}\sum_{i=1}^{n} \log((\gamma_i + \Delta)/\gamma_i). \tag{4.1}$$

At $\Delta = 0$, the model-based and model-free treatment effects agree, otherwise τ_n depends on the sample chosen via $(\gamma_i)_{i=1}^n$. In other words, the purported treatment effect absorbs features of the individuals that were present at baseline before the treatment had been applied. Consider the example of $\Delta = 2$ and two extreme cases: if all the γ_i are concentrated near zero, τ_n is large, while if all the γ_i are large, then τ_n is tiny. Thus, (3.1) converts a well-defined problem with a stable definition of treatment effect into one that is unstable with respect to sampling new individuals, and in which extrapolation of conclusions to new individuals is impossible. Although it may be argued that deviations induced by $(\gamma_i)_{i=1}^n$ should approximately average to a value close to Δ, this seems in many contexts an unreasonable assumption. R. A. Bailey (2017) writes, in her critique of a paper in *Statistical Science*:

> *I have practised as a statistician for 40 years* [...], *in no case were the experimental units a random sample from a fixed finite population. They were convenient, and were deemed to be representative enough that results on them could be extrapolated to other units.*

One final example with a similar conclusion takes the outcomes to be binary, from the logistic model

$$Y_{i0} \sim \text{Bernoulli}\left(\frac{e^{\gamma_i}}{1+e^{\gamma_i}}\right), \quad Y_{i1} \sim \text{Bernoulli}\left(\frac{e^{\gamma_i+\Delta}}{1+e^{\gamma_i+\Delta}}\right),$$

logistic regression being recovered by modelling $\gamma_i = w_i^{\mathrm{T}}\beta$. Ding and Dasgupta (2016) take as their model-free treatment effect for binary outcomes a finite-population average difference of counterfactual outcomes, the natural extension to the Hernán and Robins (2020) framework being (3.1), which is in this example

$$\tau_n = \frac{1}{n}\sum_{i=1}^{n} \frac{e^{\gamma_i}(e^{\Delta}-1)}{(1+e^{\gamma_i})(e^{\gamma_i}e^{\Delta}+1)}. \tag{4.2}$$

Except at $\tau_n = \Delta = 0$, there is a complicated dependence on the sample, as in (4.1). One could redefine τ_n as an average ratio of odds, or difference of log odds, and the correct answer would be recovered, but this undermines any motivation for a model-free definition. More generally, to define a model-free treatment effect that is stable with respect to the sample chosen, so that conclusions extrapolate to new individuals not yet observed, it seems necessary to use the structural properties of the data generating process, i.e. a model. If,

for instance, τ_n was defined as a difference in log odds ratios, and the treatment instead had an effect on the probit scale, the new definition of τ_n would again be unstable. It seems preferable, therefore, to postulate a model with a stable treatment parameter and assess the model for its compatibility with the data.

In connection with the *sharp null hypothesis* defined in §4.4, Ding and Dasgupta (2016) argued that the assumption of constant causal effect across individuals is unrealistic for binary outcomes on the scale of outcomes or expectations of outcomes, a conclusion with which we agree in view of (4.2). This was noted in support of an argument that individual units should be allowed to contribute a different treatment effect to a composite, which is permissible under (3.1); this is where our positions differ. In the case of binary outcomes, stability can only be achieved through a model that respects the $[0, 1]$ constraint on the probabilities. In the logistic case, individuals have different probabilities for the event of interest and a stable additive treatment effect on the log-odds scale. A treatment effect that differs according to individual-specific features is incorporated in the model-based definition through an interaction component, which retains the necessary stability of parameters; see McCullagh (2022, pp. 232–240).

It is clear from the above examples that the model-based and model-free definitions will give identical answers for any sample if the model belongs to a subset of the transformation family (Barndorff-Nielsen and Cox, 1994, §2.8) for which there are no other nuisance parameters needed to characterise the relevant expectation under P_γ besides γ. A restriction to the transformation family means that the group action defining the treatment effect acts on the baseline distribution P_γ via the baseline parameter space, transforming it to $P_{g\gamma}$, and group action can be described by the same group action on the observations.

4.4. **Clarification of Fisher's position.** The approach based on (3.1) is sometimes called *Neymannian* following Neyman (1923) and particularly Neyman *et al.* (1935). Within the causal inference community, another hypothesis of the form $Y_i(1) = Y_i(0)$ for all $i = 1, \ldots, n$ is sometimes contemplated. This has come to be associated with R. A. Fisher's name in the form of "Fisher's sharp null". Bailey (2017) corrected this historical misnomer, pointing out that Fisher's formulation was instead based on unit treatment additivity on an appropriate scale, as in the first two examples of §4.3. Unit treatment additivity specifies that the observation obtained by applying a particular treatment t to a particular experimental unit i is the sum of a quantity $\alpha_j + \Delta_t$ depending on the unit plus a constant reflecting the treatment applied. Unit treatment additivity is a model for the outcomes, which may be deterministic or probabilistic depending on whether α_i is treated as fixed or random. Regardless of that choice, α_i becomes random upon randomisation of units to treatments. In the fictitious idealisation used in §4.3, there was a fictitious twin for the counterfactual. No such fictitious idealisation is present in Fisher's formulation, but real twins or other blocking schemes are permitted, reflected in a block factor for twin index j and leading to $\alpha_i = \gamma_{j(i)} + \varepsilon_i$ after randomisation, where ε_i is a mean-zero and finite-variance error term.

A location-family outcome distribution is therefore implicit in the unit-treatment additivity assumption, perhaps after logarithmic transformation. These comments also resolve Holland's (1983, §6) confusion regarding D. R. Cox's position.

5. Model inadequacy

The most legitimate objection of the model-based formulation is the possibility of model misspecification. By eschewing the model, stability and interpretation is lost through the use of (3.1) or similar, but if the model is misspecified, the strategy used for inference on the notional treatment parameter may also produce misleading conclusions, and these too may be sample dependent, depending on context. This underscores the importance of assessments of model adequacy.

One source of model inadequacy is a failure to condition on relevant variables, and avoidance of systematic biases induced through such failure is the principal reason for randomisation in designed experiment. Through this route, and through the inherent temporal ordering, causation can sometimes be securely established through a suitably randomised design. More generally, Barndorff-Nielsen and Cox (1994, p. 32) illustrate the tension between relevance and generalisability through the following simple example.

> *Consider a population of individuals and an event A of interest, for instance that an individual dies of heart disease before age 70.* [...] *Now suppose that a series of new individuals is drawn randomly from the population under study and for each it is required to calculate the probability of event A* [...]. *If each probability is to be relevant to the individual in question, it must be conditional on observed relevant features, such as age, sex, smoking habits and blood pressure.* [...] *Note, however, that, especially if we condition directly, we must limit the conditioning: otherwise we would reach the position where each individual is not only unique, but also uninformative about other individuals* [...].

The question of where to limit the conditioning emerges in multiple guises throughout statistics. The above simple example is synonymous with construction of a regression model and is known in some quarters as *conditioning by model formulation*.

6. Conclusions and reflections on a burgeoning literature

Section 4.3 exemplifies how the model-free treatment effect (3.1), and other model-free definitions, are liable to change depending on the particular sample chosen, even when there is a clearly defined treatment effect that is stable over individuals. The implication is that the conclusions of inference from one study do not generalise to other populations. This observation is not new, and indeed, the ability of (3.1) to accommodate individual-specific differences is viewed as a strength by some authors. Our view, echoing Cox (1958b, 2012), is that if there is a systematic difference in treatment effect between units, then subsequent analysis or study design should seek to explain such differences. If, in a traditional modelling setting, parameters representing treatment effects were allowed to depend in an arbitrary way on the observation index, this would be in clear violation of the foundations of modelling, as formalised by McCullagh (2002).

Instability in the definition of the model-free treatment effect (3.1) or similar is the implicit reason behind several developments in the causal inference literature. For instance, Jin and Rothenhäusler (2024) proposed an approach to inference with conditional validity in finite populations, motivated by the observation that versions of τ_n for sub-populations are typically more relevant to individuals in those populations than a version specified as an average over a larger population. The need to introduce different values of τ_n for sub-populations is an inherent feature of their instability. In a similar vein, Chernozhukov *et al.* (2024) developed conditional influence functions to improve the efficiency of average treatment effect estimators conditional on a subset of covariates.

Other developments in the causal inference literature have emerged to address aspects that do not arise in the model-based formulation. Inverse probability weighting as discussed in Rosenbaum and Rubin (1983) is one way to correct for a lack of balance among the treatment groups prior to comparison by the model-free treatment effect (3.1) or generalisations thereof. Other approaches have been proposed, for instance imputation of the missing counterfactual outcomes.

In a standard regression model, there is no need for covariate classes to be balanced; maximum-likelihood inference on treatment parameters is reliable under standard regularity conditions and, while balance typically improves the precision of estimators, imbalance does not systematically bias the conclusions unless there are unmeasured confounders, a situation

that is precluded by Rosenbaum and Rubin (1983) and, to our knowledge, all discussions under the counterfactual framework. Moreover, since Fisherian inference is made either conditionally on the observed values of the covariates or on a lower dimensional ancillary statistic if one exists, extreme imbalance in the classes is reflected in the reference sets used for inference. An approach that averages over the covariate values observed for all individuals is, from a Fisherian perspective, too unconditional.

Acknowledgement. It is a pleasure to thank the referee for thoughtful observations and probing questions, and for alerting us to the differences of formulation in some of the counterfactuals literature.

Funding. H. Battey was supported by a UK Engineering and Physical Sciences Research Fellowship (EP/T01864X/1)

Supporting material. There is no supporting material associated with the paper.

Code and data availability. There is no data or code associated with the paper.

References

Bailey, R. A. (2017). Inference from randomized (factorial) experiments. *Statist. Sci.*, 32, 352–355.

Battey, H. S. and Cox, D. R. (2020). High-dimensional nuisance parameters: an example from parametric survival analysis. *Information Geometry*, 3, 119–148.

Chernozhukov, V., Newey, W. K. and Syrgkanis, V. (2024). Conditional Influence Functions. *arXiv: 2412.18080*.

Cox, D. R. (1958a). Some problems connected with statistical inference. *Ann. Math. Statist.*, 29, 357–372.

Cox, D. R. (1958b). The interpretation of the effects of non-additivity in the Latin square. *Biometrika*, 45, 69–73.

Cox, D. R. (2000). Causal inference without counterfactuals: comment. *J. Amer. Statist. Assoc.*, 95, 424–425

Cox, D. R. (2012). Statistical causality: some historical remarks. Pages 1-5 in *Causality: statistical perspectives and applications*. Edited by C. Berzuini, P. Dawid, L. Bernardinelli. Wiley, Chichester.

Dawid, A. P. (2000). Causality without counterfactuals. *J. Amer. Statist. Assoc.*, 95, 407–424.

Dawid, A. P. (2024). Potential outcomes and decision-theoretic foundations for statistical causality: Response to Richardson and Robins. *J. Causal Inference*, 12, paper number 20230058.

Ding, P. and Dasgupta, T. (2016). A potential tale of two by two tables from completely randomized experiments. *J. Amer. Statist. Assoc.*, 111, 157–168.

Fisher, R. A (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser A*, 222, 309–368.

Fisher, R. A (1935). *The Design of Experiments.* Oliver and Boyd, Edinburgh.

Hernán M. A. and Robins J. M. (2020). *Causal Inference: What If.* Chapman & Hall.

Holland, P. (1983). Statistics and causal inference. *J. Amer. Statist. Assoc.*, 81, 945–960.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, Cambridge.

Jin, Y. and Rothenhäusler, D. (2024). Tailored inference for finite populations: conditional validity and transfer across distributions. *Biometrika*, 111, 215–233.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models.* Chapman and Hall, London.

McCullagh, P. (1999). The algebraic structure of generalised linear models. University of Chicago, technical report number

McCullagh, P. (2002). What is a statistical model? *Ann. Statist.*, 30, 1225–1310.

McCullagh, P. (2022). *Ten Projects in Applied Statistics.* Springer, Cham.

Neyman, J., Iwaszkiewicz, K. and Kolodziejczyk, S. (1935). Statistical problems in agricultural experimentation (with discussion). *J. R. Statist. Soc. Suppl.*, 2, 107–180.

Popper, K. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge.* Routledge, London.

Richardson, T. S. and Robins, J. M. (2023). Potential outcome and decision theoretic foundations for statistical causality. *J. Causal Inference*, 11, paper number 20220012.

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

Robins, J., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.*, 89, 846–866.

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

Stigler, S. (1976). Discussion of 'On rereading R. A. Fisher' by L. J. Savage. *Ann. Statist.*, 4, 441–500.

Department of Mathematics, Imperial College London
Email address: h.battey@imperial.ac.uk