Causal inference and counterfactual prediction in machine learning for actionable healthcare

Mattia Prosperi[1,*], Yi Guo[2,3], Matt Sperrin[4], James S. Koopman[5], Jae S. Min[1], Xing He[2], Shannan Rich[1], Mo Wang[6], Iain E. Buchan[7], Jiang Bian[2,3]

[1]Department of Epidemiology, College of Public Health and Health Professions, College of Medicine, University of Florida, Gainesville, FL, United States of America.

[2]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, United States of America.

[3]Cancer Informatics and eHealth Core, University of Florida Health Cancer Center, Gainesville, FL, United States of America.

[4]Division of Informatics, Imaging & Data Sciences, University of Manchester, Manchester, United Kingdom.

[5]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, United States of America.

[6]Department of Management, Warrington College of Business, University of Florida, Gainesville, FL, United States of America.

[7]Institute of Population Health, University of Liverpool, Liverpool, United Kingdom.

*correspondence to: m.prosperi@ufl.edu

Abstract (208 words)

Big data, high-performance computing, and (deep) machine learning are increasingly noted as key to precision medicine –from identifying disease risks and taking preventive measures, to making diagnoses and personalising treatment for individuals. Precision medicine, however, is not only

about predicting risks and outcomes, but also about weighing interventions. Interventional clinical predictive models require the correct specification of cause and effect, and the calculation of so-called counterfactuals, i.e. alternative scenarios. In biomedical research, observational studies are commonly affected by confounding and selection bias. Without robust assumptions, often requiring *a priori* domain knowledge, causal inference is not feasible. Data-driven prediction models are often mistakenly used to draw causal effects, but neither their parameters nor their predictions necessarily have a causal interpretation. Therefore, the premise that data-driven prediction models lead to trustable decisions/interventions for precision medicine is questionable. When pursuing intervention modelling, the bio-health informatics community needs to employ causal approaches and learn causal structures. Here we discuss how target trials (algorithmic emulation of randomized studies), transportability (the license to transfer causal effects from one population to another) and prediction invariance (where a true causal model is contained in the set of all prediction models whose accuracy does not vary across different settings) are linchpins to developing and testing intervention models.

Keywords

Machine learning, artificial intelligence, data science, biostatistics, statistics, health informatics, biomedical informatics, precision medicine, prediction model, validation, causal inference, counterfactuals, transportability, prediction invariance.

Main text (4,270 manuscript)

Introduction

Advances in computing and machine learning have opened unprecedented paths to processing and inferring knowledge from big data. Deep learning has been game-changing for many analytics challenges, beating humans and other machine learning approaches in decision/action tasks, such as playing games, and aiding/augmenting tasks such as driving or recognising manipulated images. Deep learning's potential applications in healthcare have been widely speculated [1], especially in

precision medicine –the timely and tailored prevention, identification, diagnosis, and treatment of disease. However, the use of data-driven machine learning approaches to model causality, for instance, to uncover new causes of disease or assess treatment effects, carries dangers of unintended consequences. Therefore the Hippocratic principle of 'first do no harm' is being adopted [2] alongside rigorous study design, validation, and implementation, with attention to ethics and bias avoidance.

Precision medicine models are not only descriptive or predictive, e.g. assessing the mortality risk for a patient undergoing a surgical procedure, but also decision-supporting/interventional, e.g. choosing and personalising the procedure with the highest probability of favourable outcome. Predicting risks and outcomes differs from weighing interventions and intervening. Prediction calculates a future event in the absence of any action or change; intervention presumes an enacted choice that may influence the future, which requires consideration of the underlying causal structure.

Intervention imagines how the world would be if we made different choices, e.g. *"would the patient be cured if we administered amoxicillin instead of a cephalosporin for their upper respiratory tract infection?"* or *"if that pre-hypertensive patient had accomplished ten to fifteen minutes of moderate physical activity per day instead of being prescribed a diuretic, would they have become hypertensive five years later?"* By asking ourselves what would have been the effect of something if we had not taken an action, or vice versa, we are computing so-called 'counterfactuals'. Among different counterfactual options, we choose the ones that minimise harm while maximising patient benefit. A similar cognitive process happens when a deep learning machine plays a game and must decide on the next move. Such artificial neural network architecture has been fed the game ruleset, millions of game scenarios, and has learned through trial and error by playing against other human players, networks or even against itself. In each move of the game, the machine chooses the best counterfactual move based on its domain knowledge [3,4].

With domain knowledge of the variables involved in a hypothesised cause-effect route and enough data generated at random to cover all possible path configurations, it is possible to deduce causal

effects and calculate counterfactuals. Randomisation and domain knowledge are key: when either is not met, causal inference may be flawed [5].

In clinical research, randomised controlled trials (RCTs) permit direct testing of causal hypotheses since randomisation is guaranteed *a priori* by design even with limited domain knowledge. On the other hand, observational data collected retrospectively usually does not fulfil such requirements, thus limiting what secondary data analyses can discover. For instance, databases collating electronic medical records do not explicitly record domain/contextual knowledge (e.g. why one drug was prescribed over another) and are littered with many types of bias, including protopathic bias (when a therapy is given based on symptoms, yet the disease is undiagnosed), indication bias (when a risk factor appears to be associated with a health outcome, but the outcome may be caused by the reason for which the risk factor initially appeared), or selection bias (when a study population does not represent the target population, e.g. insured people or hospitalised patients).

Therefore, the development of health intervention models from observational data (no matter how big) is problematic, regardless of the method used (no matter how deep) because of the nature of the data. Fitting a machine learning model to observational data and using it for counterfactual prediction may lead to harmful consequences. One famous example is that of prediction tools for crime recidivism that convey racial discriminatory bias [6]. Any instrument inferred from existing population data may be biased by gender, sexual orientation, race, or ethnicity discrimination, and carry forward such bias when employed to aid decisions [7].

The health and biomedical informatics community, charged with maximising the utility of healthcare information, needs to be attuned to the limitations of data-driven inference of intervention models, and needs safeguards for counterfactual prediction modelling. These topics are addressed here as follows: first, we give a brief outline of causality, counterfactuals, and the problem of inferring cause-effect relations from observational data; second, we provide examples where automated

learning has failed to infer a trustworthy counterfactual model for precision medicine; third, we offer insights on methodologies for automated causal inference; finally, we describe potential approaches to validate automated causal inference methods, including transportability and prediction invariance. We aim not to criticise the use of machine learning for the development of clinical prediction models [8,9], but rather clarify that prediction and intervention models have different developmental paths and intended uses [10]. We recognise the promising applications of machine learning in healthcare for descriptive/predictive tasks rather than interventional tasks, e.g. screening images for diabetic retinopathy [11], even when diagnostic models have been shown to be susceptible to errors when applied to different population [12]. Yet, it is important to distinguish prediction works from others that seek to optimise treatment decisions and are clearly interventional [13], where validating counterfactuals becomes necessary.

Causal inference and counterfactuals

Causality has been described in various ways. For the purpose of this work, it is useful to recall the deterministic (yet can be made probabilistic) definitions by means of counterfactual conditionals – e.g. the original 1748 proposition by Hume or the 1973 formalisation by Lewis [14]– paraphrased as: an event *E* causally depends on *C* if, and only if, *E* always follows after *C* and *E* does not occur when *C* has not occurred (unless something else caused *E*). The counterfactual-based definition contains an implicit time component and works in a chained manner, where effects can become causes of other subsequent effects. Causes can be regarded as necessary, sufficient, contributory, or non-redundant [15].

Causal inference addresses the problem of ascertaining causes and effects from data. Causes can be determined through prospective experiments to observe *E* after *C* is tried or withheld, by keeping constant all other possible factors that can influence either the choice of *C* or the happening of *E,* or –what is easier and more often done– randomising the choice of *C*. Formally, we acknowledge that

the conditional probability P($E|C$) of observing E after observing C can be different from the interventional probability P($E|$do(C)) of observing E after doing C. In RCTs –when C is randomised– the 'do' is guaranteed and unconditioned, while with observational data, it is not. Causal calculus helps resolve interventional from conditional probabilities when a causal structure is assumed [16]. **Figure 1** illustrates the difference between observing and doing using biomedical target examples.

In a nutshell, the major hurdles to ascertaining causal effects from observational data include: the failure to disambiguate interventional from conditional distributions, to identify all potential sources of bias [17] and to select an appropriate functional form for all variables, i.e. model misspecification [18–20].

Two well-known types of bias are confounding and collider bias (**Figure 2**). Given an outcome, i.e. the objective of a (counterfactual) prediction, confounding occurs when there exists a variable that causes the outcome and the effect, leading to the conclusion that an exposure is associated with the outcome even though it does not cause it. For instance, cigarette smoking causes both nicotine-stained, yellow fingers and lung cancer. Yellow fingers, as the exposure or independent variable, can be spuriously associated with lung cancer if smoking, the underlying confounding variable, is unaccounted. Yellow fingers alone could be used to predict lung cancer but cleaning the skin would not reduce the risk of lung cancer. Therefore, an intervention model that used yellow fingers as the actionable item would be futile, while a model that actioned upon smoking (the cause) would be effective in reducing lung cancer risk.

A collider is a variable that is caused by both the exposure (or causes of the exposure) and the outcome (or causes of the outcome). Conditioning on a collider biases the estimate of the causal effect of exposure on outcome. A classic example involves the association between locomotor and respiratory diseases. Originally observed in hospitalised patients and thought biologically plausible,

this association could not be established in the general population [21]. In this case, hospitalisation status functions as a collider because it introduces selection bias, as people with locomotor disease or respiratory disease have higher risk of being admitted to hospital.

Another example of collider bias is the obesity paradox [22]. This paradox refers to the counterintuitive evidence of lower mortality among people who are obese within certain clinical subpopulations, e.g. in patients with heart failure. A more careful consideration of the covariate-outcome relationship in this case reveals heart failure is a collider. Had an intervention been developed by means of such model, treating obesity would not have been suggested as an actionable feature to reduce the risk of mortality.

Causal inference can become more complex when a variable may be mistaken for a confounder but actually functions as a collider. This phenomenon is called M-bias since the associated causal diagram is usually drawn in an M-shaped form [23,24]. A classic M-bias example is the effect of education on diabetes, controlled through family history of diabetes and income. In a hypothetical study, it could be reasonable to regard mother's (or father's) history of diabetes as a confounder, because it is associated with both education level and diabetes status, and it is not caused by either. However, family history of diabetes' associations with the education and diabetes are not causal but are in turn confounded by family income and family genetic risk for diabetes, respectively, that might not be measured as input (**Figure 3**). At this point, mother's diabetes becomes a collider, and including it would induce a biased association between education and diabetes through the links from family income and genetic risk. Specifically, the estimate of the causal effect of education on diabetes would be biased. In general, including mother's diabetes in the input covariate would lead to bias both if there was a zero or non-zero causal effect [25]; however, if the unmeasured covariates were included, the bias would be resolved (by a so-called backdoor path blocking) [26]. The M-bias

example shows how the causal structure choice (which could be machine learned) can influence the causal effect inference; we will discuss the two more in detail later a specific section.

For brevity, we do not describe moderators, mediators, and other important concepts in causality modelling. Nonetheless, it is useful to mention instrumental variables, which determine variation in an explanatory variable, e.g. a treatment, but have no independent effect on the outcome of interest. Instrumental variables, therefore, can be used to resolve unmeasured confounding in absence of prospective randomisation.

An old neural network fiasco and a new possible paradox

In 1997, Cooper et al. [27] investigated several machine learning models, including rule-based and neural networks, for predicting mortality of hospital patients admitted with pneumonia. The neural network greatly outperformed logistic regression; however, the authors discouraged using black box models in clinical practice. They showed how the rule-based method learned that 'IF patient admitted (with pneumonia) has history of asthma THEN patient has lower risk of death from pneumonia' [28]. This counterintuitive association was also later confirmed using generalised additive models [29]. The physicians explained that patients admitted with pneumonia and known history of asthma would likely be transferred to intensive care and treated aggressively, thus having higher odds of survival than the general population admitted with pneumonia. The authors recommended to employ interpretable models instead of black boxes, to identify counterintuitive, surprising patterns and remove them. At this point, the model development is no longer automated and requires domain knowledge. Upon reflection, those models inferred without modifications, either interpretable or black box, would have worked well at predicting mortality but they could not have been used to test new interventions to reduce mortality, as the recommended actions would have consequentially led to 'less care' for asthmatic patients.

More recently, a possible data-driven improvement in the evaluation of fall risk in hospitals was investigated [30]. Standard practice involves a nurse-led evaluation of patients' history of falls, comorbidities, mental health, gait, ambulatory aids, and intravenous therapy summarised with the Morse scale. To assess the predictive ability of the Morse scale (standard practice), its individual components, and new expert-based predictors (e.g. extended clinical profiles and information on hospital staffing levels), a matched study was performed including patients with and without a fall. Logistic regression and decision trees were used. The additional variables hypothesised by the experts were associated with the outcome and all new models yielded higher discrimination than the Morse scale, but a surprising finding was observed: in all configurations, older patients were at a lower risk of falls. This is contrary to current expert's knowledge, which associates older age with increased frailty and therefore fall risk. If such model were used for intervention, it would not prioritise the elderly for fall prevention –a potentially devastating consequence of data-driven inference. It is uncertain if this old age paradox is due to a bias. One possible explanation is that older patients are usually monitored and aided more frequently because they are indeed at higher risk, while younger people may be more independent and less prone to accept assistance. Other issues at play could be survivorship bias, selectively unreported falls, and study design. One possible approach to bias reduction is to design the study and extract the data by simulating an RCT, where causal effects on "randomised" interventions can be estimated directly, as we discuss in the next section.

The target trial

Target trials refer to RCTs that can be emulated using data from large, observational databases to answer causal questions of comparative treatment effect [31]. Although RCTs are the gold standard for discerning causal effects, there exists many scenarios in which they are neither feasible nor ethical to conduct. Alternatively, observational data appropriately adjusted for measured confounding bias –

for instance, via propensity score matching– can be used to emulate randomised treatment assignment; this may be feasible with electronic medical records where many individual-level attributes can be linked to resolve bias. The target trial protocol requires prospective enrolment-like eligibility criteria, a description of treatment strategies and assignment procedures, the identification of time course from a baseline to the outcome, a causal query (e.g. treatment effect), and an analysis plan (e.g. a regression model), as shown in **Table 1**.

As an example of the target trial framework, data from public surveillance and clinical claims repositories were used to replicate two RCTs, one investigating treatment effects on colorectal cancer and the other on pancreatic adenocarcinoma [32]. Each study explicitly adhered to the target trial framework, deviating from the RCT design only in the assignment procedures, justifiably due to lack of randomisation. The results were consistent with the target trials –all of which reported a null effect. In contrast, when the authors modelled the treatment effects using a non-RCT-like study design with the same variables, the mortality estimates were both inconsistent with the target trials. These examples demonstrate the need to uphold target trial design in the investigation of treatment effects using observational data. Moreover, coupled with machine learning methods equipped to extrapolate more useful information from big data sources, the target trial framework has the potential to serve as the foundation for exploring causal processes currently unknown.

Causal effect inference and automated learning of causal structures

Prediction models inferred automatically using data without any domain knowledge –from linear regression to deep learning– only approximate a function, and do not necessarily hold a causal meaning. Instead, by estimating interventional in place of conditional probabilities, models can reproduce causal mechanisms. Through counterfactual predictions, models become interventional, actionable, and avoid the pitfalls such as those described in the pneumonia and fall risk examples. In the previous sections we showed that it is possible to directly estimate causal effects by generating

239 data through RCTs or by simulating RCTs with observational data. Here, we delve further into the
240 approaches to unveiling and disambiguating causality from observational data, including the
241 assumptions to be made. We can categorise two main tasks: (i) estimating causal effects, and (ii)
242 learning causal structures. In the first one, a causal structure, or a set of cause-effect relationships
243 and pathways, is defined a priori, input variables are fixed, and causal effects are estimated for a
244 specific input-output pair, e.g. causal effect of diabetes on glaucoma. Directed acyclic graphs
245 (DAGs) [33] –also known as Bayesian networks– and structural equation models [34,35] are often used to
246 model such structures. While with RCT data the estimation of causal effects can be done directly,
247 the estimation of causal effects from observational data requires thoughtful handling of potential
248 biases and confounding. Methods like inverse probability weighting attempt to weigh single
249 observations to mimic the effects of randomisation with respect to one variable of interest (e.g. an
250 exposure or a treatment) [36]. Other techniques include targeted maximum likelihood estimation [37–39],
251 g-methods [40,41], and propensity score matching [42]. Often, these estimators can be coupled with
252 machine learning, e.g. causal decision trees [43], Bayesian regression trees [44], and random forests for
253 estimating individual treatment effects [45]. As previously mentioned, model misspecification –i.e.
254 defining the wrong causal structure and the choice of variables to handle confounding and bias– can
255 lead to wrong estimation of causal effects. With big data, especially datasets with numerous features,
256 choosing adjustments and even over-adjusting using all variables, is problematic. Feature selection
257 algorithms based on conditional independence scoring have been proposed [46].
258 Automated causal structure learning uses conditional independence tests and structure search
259 algorithms over given DAGs subject to certain assumptions, e.g. 'causal sufficiency' that is no
260 unmeasured common causes and no selection bias. In 1990, an important result on independence
261 and conditional independence constraints –the d-separation equivalence theorem [47]– led to the
262 development of automated search and ambiguity resolution of causal structures from data, through

so-called patterns and partial ancestral graphs. When assumptions are met (Markov/causal faithfulness [48]), there are asymptotically correct procedures that can predict an effect or raise an ambiguity, and determine graph equivalence [49]. However, the probability of an effect cannot be obtained without deciding on a prior distribution of the graphs and parameters. Also, the number of graphs is super-exponential in the number of observed variables and may even be infinite with hidden variables [50], making an exhaustive search computationally unfeasible [51]. Today, several heuristic methods for causal structure search are available, from the PC algorithm that assumes causal sufficiency, to others like FCI or RFCI that extend to latent variables [52,53].

The enrichment of deep learning with causal methods also provides interesting new insights to address bias. For instance, a theoretical analysis for identification of individual treatment effect under strong ignorability has been derived [54], and an approach to exploit instrumental variables for counterfactual prediction within deep learning is also available [55].

Model transportability and prediction invariance

Validation of causal effects under determined causal structures is especially needed when such effects are estimated in limited settings, e.g. RCTs. Transportability is a data fusion framework for external validation of intervention models and counterfactual queries. As defined by Pearl and Bareinboim [56], transportability is a "license to transfer causal effects learned in experimental studies to a new population, in which only observational studies can be conducted." By combining datasets generated under heterogeneous conditions, transportability provides formal mathematical tools to (i) evaluate whether results from one study (e.g. a causal relationship identified in an RCT) could be used to generate valid estimates of the same causal effect in another study of different setting (e.g. an observational study of the same causal effect in a different population); and (ii) estimate what the causal effect would have been if the study had been conducted in the new setting [57,58]. The framework utilises selection diagrams [59], encoding the causal relationships of variables of interest in a

study population, and about the characteristics in which the target and study populations differ. If the structural constraints among variables in the selection diagrams are resolvable through the do-calculus, a valid estimate of the causal effect in the target population can be calculated using the extant causal effect from the original study, which means that the observed causal effect is transportable.

One of Pearl's transportability examples is shown in **Figure 4**. In this example, a RCT is conducted in city *A* (original environment) and a causal effect of treatment *x* on outcome *y*, $P(y|do(x))$, is determined. We wish to generalise if the treatment works also in the population of city *B* (target environment) where only observational data is available, since it happens that the age distribution in city *A*, $P(z)$, is different than that $P^*(z)$ in city *B*. The city *B* specific *x*-to-*y* causal effect $P^*\left(y\middle|do(x)\right)$ is estimated as:

$$P^*\left(y\middle|do(x)\right) = \sum_z P(y|do(x),z)P^*(z)$$

In this transport formula, the age-specific causal effects estimated in the RCT, $P(y|do(x),z)$, is combined with the observed age distribution in the target population, $P^*(z)$, to obtain the causal effect $P^*\left(y\middle|do(x)\right)$ in city *B*.

On the other hand, a causal effect is not always transportable. Following the example above, the *x*-to-*y* causal effect is not transportable from City *A* to City *B* if only the overall causal effect $P(y|do(x))$ is known whereas the age-specific causal effect $P(y|do(x),z)$ is unknown.

Transportability theory is being extended to a variety of more complex causal relationships, e.g. sample selection bias [58], leaping forward from toy examples to real-world problems [60]. Therefore – linking back with the problematic examples we discussed in the previous sections– one could use transportability to determine how the asthma or old age effects are/are not transportable from one population to another. It is also worth noting how transportability evokes the field of domain

adaptation, which aims to learn a model in one source population that can be used in a different target distribution. In fact, domain adaptation has been employed to address sample selection bias [61]. An interesting next of kin to transportability is prediction invariance [62]. Among all models that show invariance in their predictive accuracy across different experimental settings and interventions, there is a high probability that the causal model will be a member of that set. For example, Schulam and Saria [72] introduced the counterfactual Gaussian process to predict continuous-time trajectories under irregular sampling, handling biases arising from clinical protocols. In another work, aimed at addressing issues of supervised learning when training and target distributions differ (i.e. dataset shift), Saria *et al.* [63] proposed the 'surgery estimator', defined as an interventional distribution [16] that is invariant to differences across environments. The surgery estimator works by learning a relationship in the training data that is generalisable to the target population, by incorporating prior knowledge about the data generating process that are expected to differ between the original and target populations. It was applied in real-world cases where causal structures were unknown.

Conclusions

We explored common pitfalls of data-driven developments in machine learning for healthcare, distinguishing between prediction and intervention models that are actionable in support of clinical decision-making. Importantly, the development of intervention models requires careful consideration of causality. Hernan et al. [64] commented that "a recent influx of data analysts, many not formally trained in statistical theory, bring a fresh attitude that does not *a priori* exclude causal questions" yet called –and we strongly endorse such call– for training curricula in data science that well-differentiate descriptive, prediction, and intervention modelling.

Undertaking causal machine learning is key to ethical artificial intelligence for healthcare, equivalent to a doctor's oath to "first do no harm" [65]. Healthcare intervention models involve actionable inputs and need –implicitly or explicitly– to model causal pathways to compute the correct counterfactuals.

There are ongoing discussions in the machine learning community about model explainability for bias avoidance and fairness in decisions [66]. Bias is a core topic in causal theory. Explainability may be a 'weaker' model property than causality. Explaining the role of input variables in changing the output of a black box neither assures a correct interpretation of the input-output mechanism nor unveils the cause-effect relationships. For instance, in a deep learning system that predicts the risk of heart attack, a subsequent analysis could be able to explain that the input variables 'race' and 'blood pressure' affect the risk, but could not say if these findings are causal, since they may be biased by stratification, unmeasured confounders, or mediated by other factors in the causal pathway. Fairness in machine learning aims at developing models that avoid social discrimination due to historically biased data and involves the same conceptual hurdles as learning from observational data. In fact, the usage of causal models has been advocated to identify and mitigate discriminatory relationships in data [67]. Recently, a study in cancer prognostics presented a causal structure coupled to deep learning to eliminate collider bias and provide unbiased individual predictions [68], although it did not explicitly test for transportability.

For context-specific intervention models, where a causal structure is available or a target trial design can be devised, we then recommend evaluation of model transportability for a given set of action queries, e.g. treatment options or risk modifiers. For broader exploratory analyses where causal structures need to be identified or clarified, prediction invariance could be used. Transportability and prediction invariance could become core tools to reporting protocols for intervention models, in line with the current standards for prognostic and diagnostic models [69]. A transportable model can be integrated into clinical guidelines to augment healthcare with action-savvy predictions, in pursuit of better precision medicine.

Acknowledgments

Dr. Bian's, Guo's and Prosperi's research for this work was in part supported by University of Florida's Creating the Healthiest Generation - Moonshot initiative, supported by the UF Office of the Provost, UF Office of Research, UF Health, UF College of Medicine and UF Clinical and Translational Science Institute. Dr. Wang's work on this research was supported in part by the Lanzillotti-McKethan Eminent Scholar Endowment.

Author Contributions

MP, YG, MS, JB: conceived and designed the experiments, contributed materials/analysis tools, wrote the paper.

JK, IB: conceived and designed the experiments, wrote the paper.

JM XE, SR, MW: contributed materials/analysis tools, wrote the paper.

Competing Interests Statement

The authors have no competing interests as defined by Nature Research, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

References

1. Norgeot, B., Glicksberg, B. S. & Butte, A. J. A call for deep-learning healthcare. *Nature Medicine* (2019). doi:10.1038/s41591-018-0320-3

2. Wiens, J. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* (2019). doi:10.1038/s41591-019-0548-6

3. Silver, D. *et al.* Mastering the game of Go without human knowledge. *Nature* (2017). doi:10.1038/nature24270

4. Jin, P., Keutzer, K. & Levine, S. Regret minimization for partially observable deep

reinforcement learning. in *35th International Conference on Machine Learning, ICML 2018* (2018).

5. Pearl, J. & Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*. (Basic Books, Inc., 2018).

6. Chouldechova, A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* (2017). doi:10.1089/big.2016.0047

7. Kusner, M., Loftus, J., Russell, C. & Silva, R. Counterfactual fairness. in *Advances in Neural Information Processing Systems* (2017).

8. Christodoulou, E. *et al.* A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology* (2019). doi:10.1016/j.jclinepi.2019.02.004

9. Bian, J., Buchan, I., Guo, Y. & Prosperi, M. Statistical thinking, machine learning. *J. Clin. Epidemiol.* (2019). doi:10.1016/j.jclinepi.2019.08.003

10. Baker, R. E., Peña, J. M., Jayamohan, J. & Jérusalem, A. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biol. Lett.* (2018). doi:10.1098/rsbl.2017.0660

11. Gulshan, V. *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA - J. Am. Med. Assoc.* (2016). doi:10.1001/jama.2016.17216

12. Winkler, J. K. *et al.* Association between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatology* (2019). doi:10.1001/jamadermatol.2019.1735

13. Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C. & Faisal, A. A. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med.* (2018). doi:10.1038/s41591-018-0213-5

14. Lewis, D. K. Causation. *J. Philos.* **70**, (1973).

15. Mackie, J. L. *The Cement of the Universe.* (Oxford, Clarendon Press, 1974).

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

17. Rothman, K. J., Greenland, S. & Lash, T. *Modern Epidemiology, 3rd Revised Edition. Lippincott Williams & Williams* (2012). doi:10.1017/CBO9781107415324.004

18. Lehmann, E. L. Model Specification: The views of Fisher and Neyman, and later developments. *Stat. Sci.* (1990). doi:10.1214/ss/1177012164

19. Vansteelandt, S., Bekaert, M. & Claeskens, G. On model selection and model misspecification in causal inference. *Stat. Methods Med. Res.* (2012). doi:10.1177/0962280210387717

20. Asteriou, D., Hall, S. G., Asteriou, D. & Hall, S. G. Misspecification: Wrong Regressors, Measurement Errors and Wrong Functional Forms. in *Applied Econometrics* (2016). doi:10.1057/978-1-137-41547-9_8

21. Sackett, D. L. Bias in analytic research. *J. Chronic Dis.* (1979). doi:10.1016/0021-9681(79)90012-2

22. Banack, H. R. & Kaufman, J. S. The 'obesity paradox' explained. *Epidemiology* (2013). doi:10.1097/EDE.0b013e31828c776c

23. Pearl, J. Causal diagrams for empirical research. *Biometrika* (1995). doi:10.1093/biomet/82.4.669

24. Greenland, S., Pearl, J. & Robins, J. M. Causal diagrams for epidemiologic research. *Epidemiology* (1999). doi:10.1097/00001648-199901000-00008

25. Westreich, D. & Greenland, S. The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology* (2013). doi:10.1093/aje/kws412

26. Wei, L., Brookhart, M. A., Schneeweiss, S., Mi, X. & Setoguchi, S. Implications of m bias in epidemiologic studies: A simulation study. *Am. J. Epidemiol.* (2012). doi:10.1093/aje/kws165

27. Cooper, G. F. *et al.* An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif. Intell. Med.* (1997). doi:10.1016/S0933-3657(96)00367-3

28. Ambrosino, R., Buchanan, B. G., Cooper, G. F. & Fine, M. J. The use of misclassification costs to learn rule-based decision support models for cost-effective hospital admission strategies. *Proc. Annu. Symp. Comput. Appl. Med. Care* (1995).

29. Caruana, R. *et al.* Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015). doi:10.1145/2783258.2788613

30. Lucero, R. J. *et al.* A data-driven and practice-based approach to identify risk factors associated with hospital-acquired falls: Applying manual and semi- and fully-automated methods. *Int. J. Med. Inform.* (2019). doi:10.1016/j.ijmedinf.2018.11.006

31. Hernán, M. A. & Robins, J. M. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am. J. Epidemiol.* (2016). doi:10.1093/aje/kwv254

32. Petito, L. C. *et al.* Estimates of Overall Survival in Patients With Cancer Receiving Different Treatment Regimens: Emulating Hypothetical Target Trials in the Surveillance, Epidemiology, and End Results (SEER)–Medicare Linked Database. *JAMA Netw. Open* **3**, e200452–e200452 (2020).

33. Pearl, J. Causal diagrams for empirical research. *Biometrika* (1995). doi:10.1093/biomet/82.4.669

34. Westland, J. C. An introduction to structural equation models. in *Studies in Systems, Decision and Control* (2019). doi:10.1007/978-3-030-12508-0_1

35. Bollen, K. A. & Pearl, J. Eight Myths About Causality and Structural Equation Models. in (2013). doi:10.1007/978-94-007-6094-3_15

36. Hernán, M. A. & Robins, J. M. Estimating causal effects from epidemiological data. *Journal of*

Epidemiology and Community Health (2006). doi:10.1136/jech.2004.029496

37. van der Laan, M. J. & Rubin, D. Targeted maximum likelihood learning. *Int. J. Biostat.* (2006). doi:10.2202/1557-4679.1043

38. Schuler, M. S. & Rose, S. Targeted maximum likelihood estimation for causal inference in observational studies. *Am. J. Epidemiol.* (2017). doi:10.1093/aje/kww165

39. Rose, S. & van der Laan, M. J. Targeted Learning: Causal Inference for Observational and Experimental Data. *Target. Learn. Causal Inference Obs. Exp. Data* (2011). doi:10.1007/978-1-4419-9782-1

40. Naimi, A. I., Cole, S. R. & Kennedy, E. H. An introduction to g methods. *Int. J. Epidemiol.* (2017). doi:10.1093/ije/dyw323

41. Robins, J. M. & Hernán, M. A. Estimation of the causal effects of time-varying exposures. in *Longitudinal Data Analysis* (2008).

42. Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* (1983). doi:10.1093/biomet/70.1.41

43. Li, J., Ma, S., Le, T., Liu, L. & Liu, J. Causal Decision Trees. *IEEE Trans. Knowl. Data Eng.* (2017). doi:10.1109/TKDE.2016.2619350

44. Hahn, P. R., Murray, J. & Carvalho, C. M. *Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects*. *SSRN* (2017). doi:10.2139/ssrn.3048177

45. Lu, M., Sadiq, S., Feaster, D. J. & Ishwaran, H. Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods. *J. Comput. Graph. Stat.* (2018). doi:10.1080/10618600.2017.1356325

46. Schneeweiss, S. *et al.* High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* (2009). doi:10.1097/EDE.0b013e3181a663cc

47. Verma, T. & Pearl, J. Causal Networks: Semantics and Expressiveness. in *Machine Intelligence and Pattern Recognition* (1990). doi:10.1016/B978-0-444-88650-7.50011-1

48. Jaber, A., Zhang, J. & Bareinboim, E. Causal identification under Markov equivalence. in *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018* (2018).

49. Richardson, T. Equivalence in Non-Recursive Structural Equation Models. in *Compstat* (1994). doi:10.1007/978-3-642-52463-9_59

50. Heckerman, D., Meek, C. & Cooper, G. F. A Bayesian approach to causal discovery. *Studies in Fuzziness and Soft Computing* (2006). doi:10.1007/10985687_1

51. Peter Spirtes, C. G. and R. S. Causation, Prediction, and Search. 2nd edn. MIT Press. *Stat. Med.* (2003). doi:10.1002/sim.1415

52. Glymour, C., Zhang, K. & Spirtes, P. Review of causal discovery methods based on graphical models. *Front. Genet.* (2019). doi:10.3389/fgene.2019.00524

53. Colombo, D. & Maathuis, M. H. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* (2015).

54. Shalit, U., Johansson, F. D. & Sontag, D. Estimating individual treatment effect: Generalization bounds and algorithms. in *34th International Conference on Machine Learning, ICML 2017* (2017).

55. Hartford, J., Lewis, G., Leyton-Brown, K. & Taddy, M. Deep {IV}: A Flexible Approach for Counterfactual Prediction. in *Proceedings of the 34th International Conference on Machine Learning* (eds. Precup, D. & Teh, Y. W.) **70**, 1414–1423 (PMLR, 2017).

56. Pearl, J. & Bareinboim, E. External validity: From do-calculus to transportability across populations. *Stat. Sci.* (2014). doi:10.1214/14-STS486

57. Dahabreh, I. J., Robertson, S. E., Tchetgen, E. J., Stuart, E. A. & Hernán, M. A. Generalizing causal inferences from individuals in randomized trials to all trial□eligible individuals.

Biometrics (2019). doi:10.1111/biom.13009

58. Bareinboim, E. & Pearl, J. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences of the United States of America* (2016). doi:10.1073/pnas.1510507113

59. Pearl, J. & Bareinboim, E. Transportability of causal and statistical relations: A formal approach. in *Proceedings - IEEE International Conference on Data Mining, ICDM* (2011). doi:10.1109/ICDMW.2011.169

60. Lee, S., Correa, J. D. & Bareinboim, E. General identifiability with arbitrary surrogate experiments. in *35th Conference on Uncertainty in Artificial Intelligence, UAI 2019* (2019).

61. Huang, J., Smola, A. J., Gretton, A., Borgwardt, K. M. & Schölkopf, B. Correcting sample selection bias by unlabeled data. in *Advances in Neural Information Processing Systems* (2007). doi:10.7551/mitpress/7503.003.0080

62. Peters, J., Bühlmann, P. & Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *J. R. Stat. Soc. Ser. B Stat. Methodol.* (2016). doi:10.1111/rssb.12167

63. Subbaswamy, A., Schulam, P. & Saria, S. Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport. BT - The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan. 3118–3127 (2019).

64. Hernán, M. A., Hsu, J. & Healy, B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *CHANCE* **32**, 42–49 (2019).

65. Wiens, J. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* (2019). doi:10.1038/s41591-019-0548-6

66. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* (2019). doi:10.1038/s42256-019-0048-x

67. Kusner, M. J. & Loftus, J. R. The long road to fairer algorithms. *Nature* (2020). doi:10.1038/d41586-020-00274-3

68. van Amsterdam, W. A. C., Verhoeff, J. J. C., de Jong, P. A., Leiner, T. & Eijkemans, M. J. C. Eliminating biasing signals in lung cancer images for prognosis predictions with deep learning. *npj Digit. Med.* (2019). doi:10.1038/s41746-019-0194-x

69. Moons, K. G. M. *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann. Intern. Med.* (2015). doi:10.7326/M14-0698

Figure Legends

Figure 1. *Conditional vs. interventional probabilities.* When we observe data, e.g. electronic medical records, we can learn a model that predicts the probability of a disease D given certain risk factors R, i.e. $P(D|R)$, or a model that predicts the chance of an health outcome O for a given treatment T, i.e. $P(O|T)$. However, these models cannot be used to support decisions, because they assume that variables of the model remain unchanged, people keep their lifestyles, and the standard of care is followed. When a risk factor is modified or a new treatment is tested, e.g. in a randomised controlled trial, then we 'make' new data, and compute different probabilities, which are $P(D|\mathrm{do}(R))$ and $P(O|\mathrm{do}(T))$. Conditional and interventional probabilities are not necessarily the same, e.g. treatments are randomised in trials, while they are not in clinical practice.

Figure 2. *Examples of confounding bias and collider bias.* Confounding (panel a) can occur when there exists a common cause for both exposure and outcome, while a collider (panel b) is a common effect of both exposure and outcome. Not including a confounder or including

549 a collider in a model results in biased associations.

550 **Figure 3.** ***An example of M-bias.*** When estimating the effect of education level on diabetes risk,
551 mother's history of diabetes could be mistaken as a confounder and included in a model, but
552 it is a collider by the effect of history of family income and genetic risk.

553 **Figure 4.** ***A selection diagram for illustrating transportability.*** A causal effect of treatment *x* on
554 outcome *y*, $P(y|do(x))$, is found through an RCT, and quantified in the original environment of
555 city *A* (panel a). The *x*-to-*y* causal effect is transportable from City *A* to City *B* as $P^*(y|do(x))$
556 (panel b) if both the overall causal effect $P(y|do(x))$ and the age-specific causal effect
557 $P(y|do(x), z)$ are known, whilst it is not transportable if the latter is unknown.

558

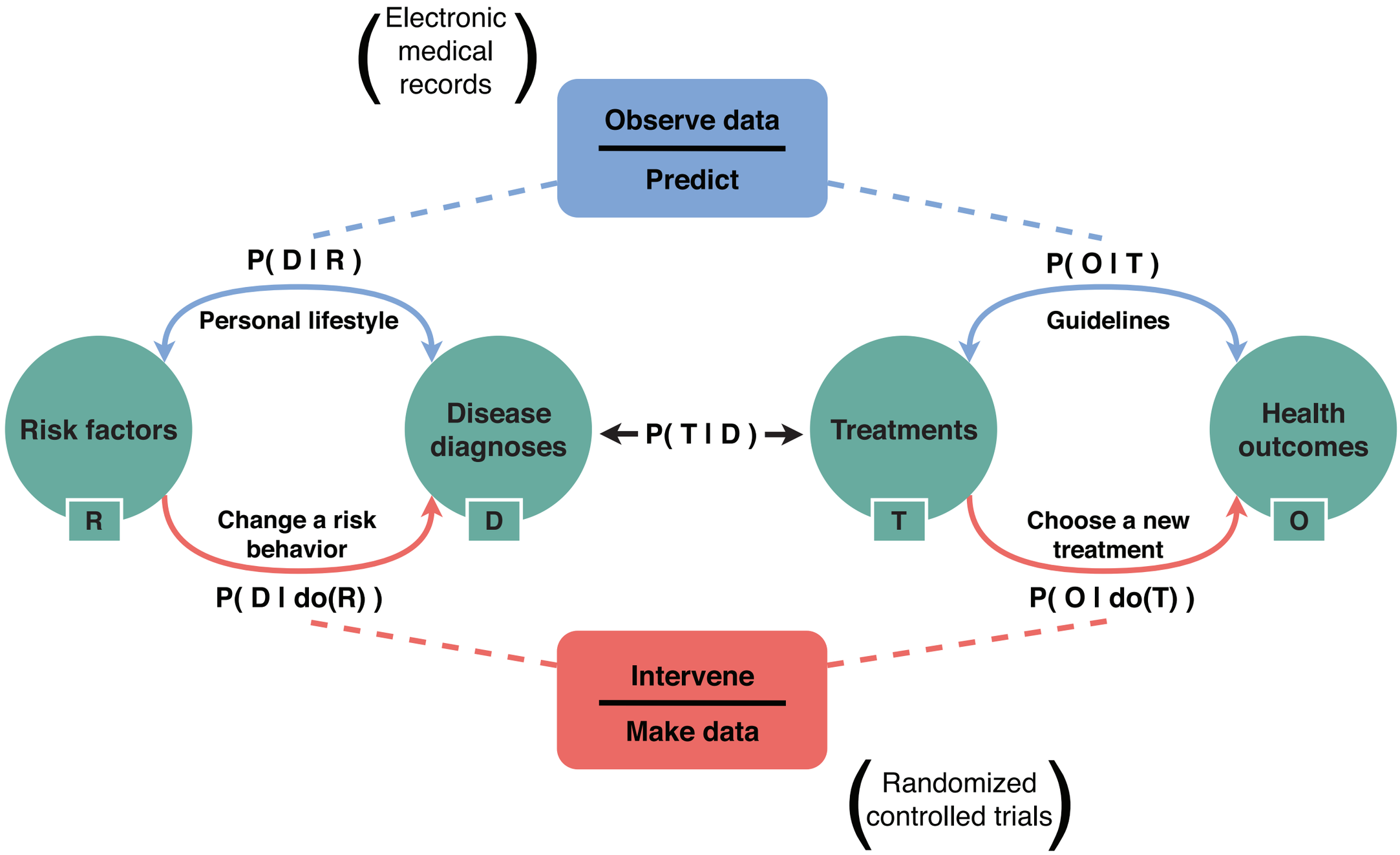
559 **Tables**

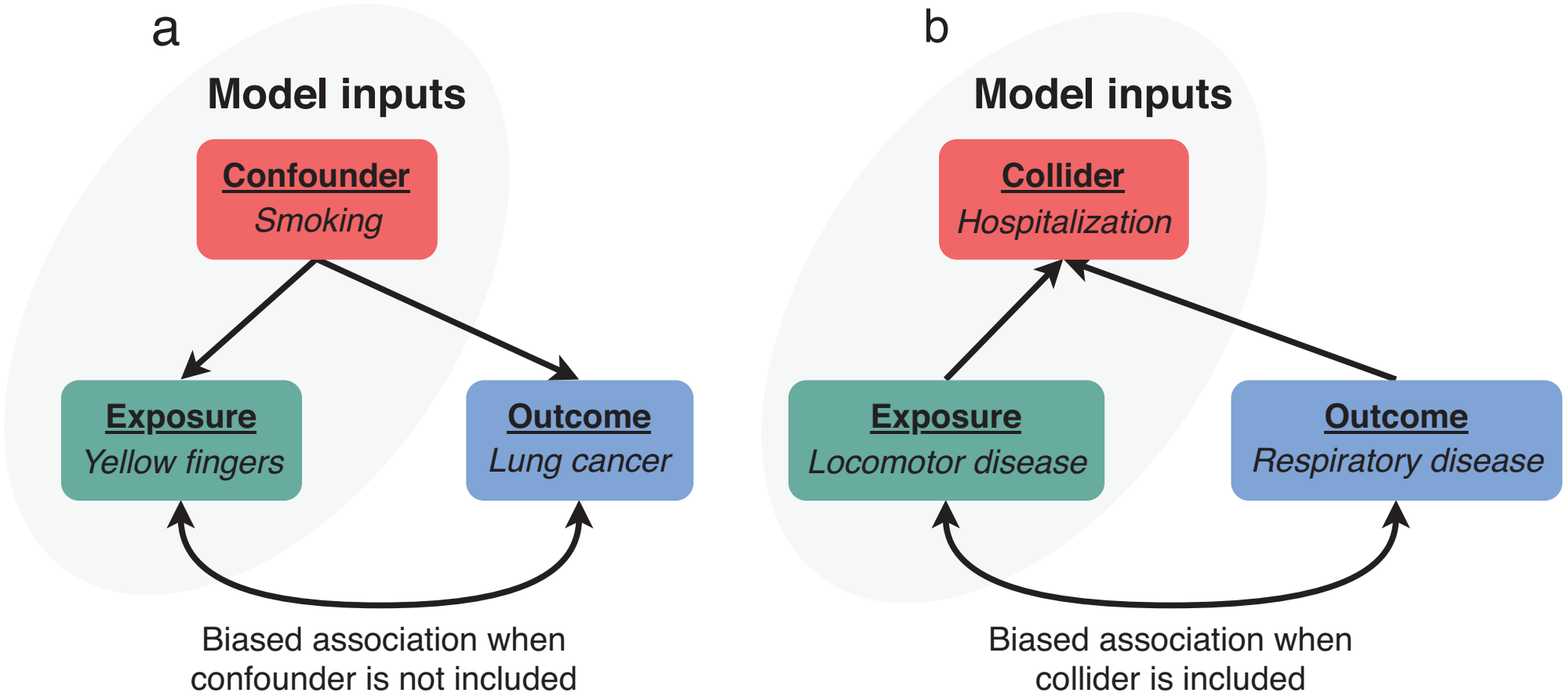
560 **Table 1.** ***The target trial protocol.*** Emulation of a randomised clinical trial using observational
561 data and algorithmic randomisation, with the objective to reduce bias and allow more reliable
562 treatment effects estimates.

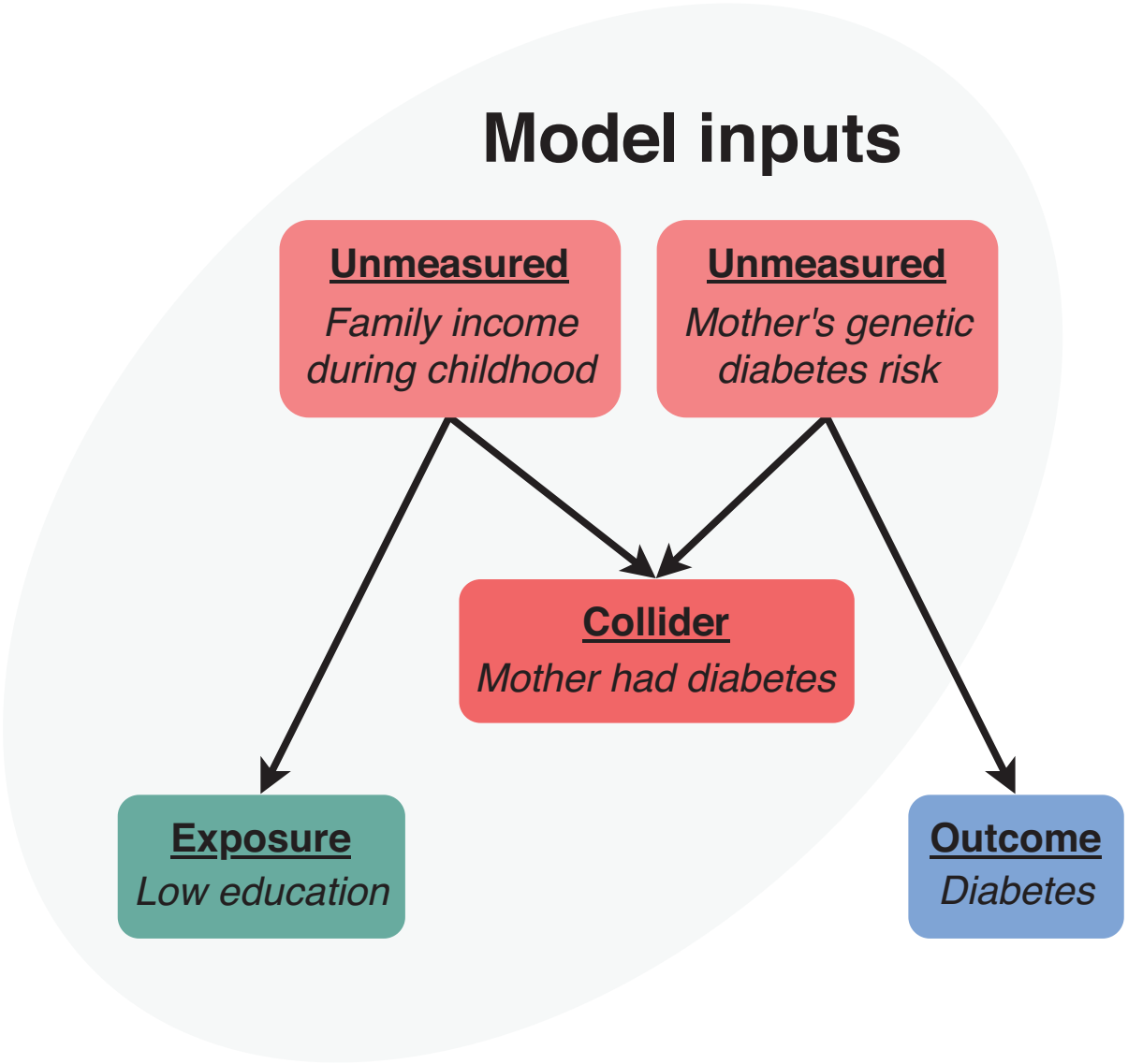
	RCT	Target Trial
Data source	Prospective	Observational
Sample size	Small	Large
Variables	Few	Many
Eligibility and time zero (baseline)	Straightforward	Problematic (e.g. multiple baseline points, follow up requirements)
Treatment assignment	Randomised by design	Randomised algorithmically (e.g. propensity score

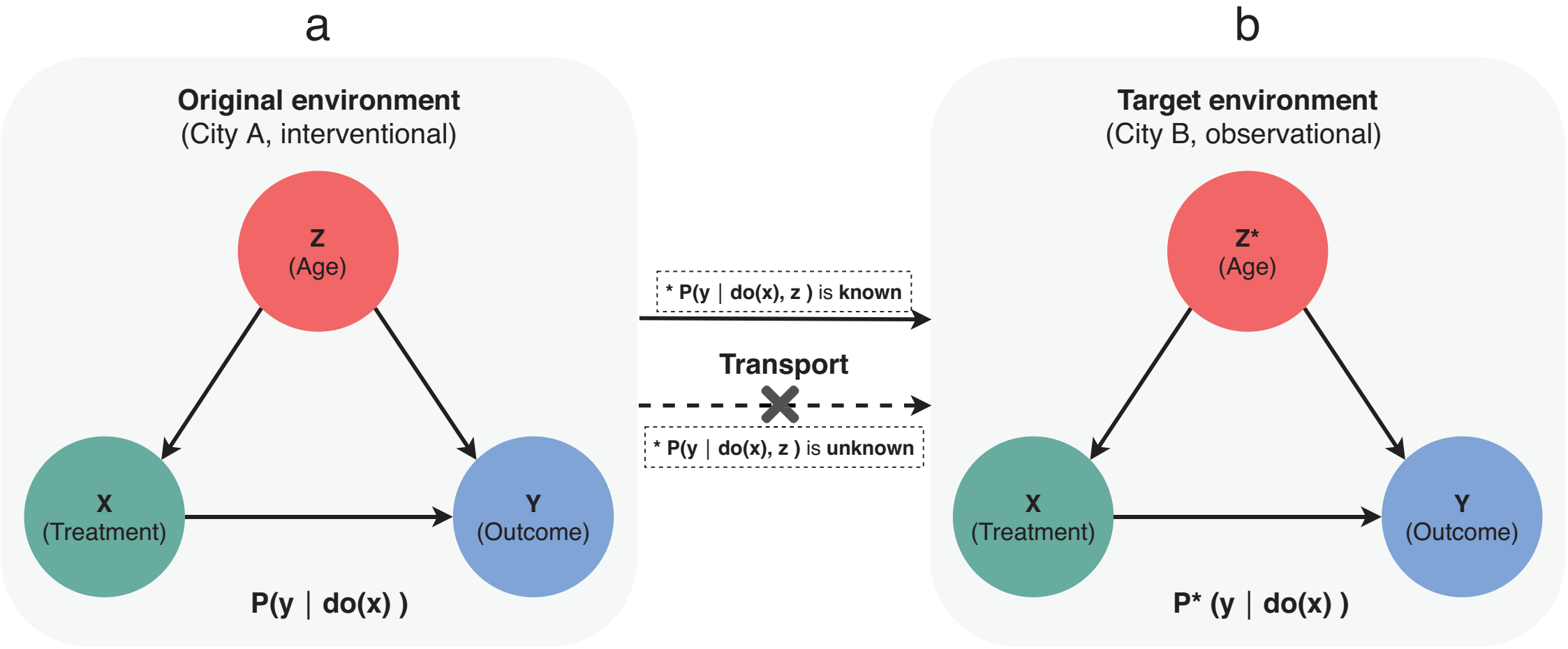
		matching)
Outcome evaluation	Flexible	Flexible (with some caveats for blind outcome studies)
Analysis plan	Relatively straightforward (e.g. intention to treat) and flexible (e.g. Bayesian adaptive), but can further require bias correction (e.g. g-formula)	More complex (need also to model treatment assignment) yet can use same techniques as for RCT (e.g. g-formula)
Risk of bias	Relatively low	Possible (e.g. residual confounding, wrong choice of time zero)
Flexibility to assess extra-protocol causal effects	Limited	High

563


Electronic medical records
Observe data
Predict
P(D | R)
Personal lifestyle
P(O | T)
Guidelines
Risk factors
R
Disease diagnoses
D
P(T | D)
Treatments
T
Health outcomes
O
Change a risk behavior
P(D | do(R))
Choose a new treatment
P(O | do(T))
Intervene
Make data
Randomized controlled trials


a
Model inputs
Confounder
Smoking
Exposure
Yellow fingers
Outcome
Lung cancer
Biased association when
confounder is not included
b
Model inputs
Collider
Hospitalization
Exposure
Locomotor disease
Outcome
Respiratory disease
Biased association when
collider is included


Model inputs
Unmeasured
Family income during childhood
Unmeasured
Mother's genetic diabetes risk
Collider
Mother had diabetes
Exposure
Low education
Outcome
Diabetes


a
b
Original environment
(City A, interventional)
Z
(Age)
X
(Treatment)
Y
(Outcome)
P(y | do(x))
* P(y | do(x), z) is known
Transport
* P(y | do(x), z) is unknown
Target environment
(City B, observational)
Z*
(Age)
X
(Treatment)
Y
(Outcome)
P* (y | do(x))