

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

38 **Keywords**

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

42 **Main text** (4,270 manuscript)

43 *Introduction*

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

49 precision medicine –the timely and tailored prevention, identification, diagnosis, and treatment of
50 disease. However, the use of data-driven machine learning approaches to model causality, for
51 instance, to uncover new causes of disease or assess treatment effects, carries dangers of unintended
52 consequences. Therefore the Hippocratic principle of ‘first do no harm’ is being adopted ² alongside
53 rigorous study design, validation, and implementation, with attention to ethics and bias avoidance.
54 Precision medicine models are not only descriptive or predictive, e.g. assessing the mortality risk for
55 a patient undergoing a surgical procedure, but also decision-supporting/interventional, e.g. choosing
56 and personalising the procedure with the highest probability of favourable outcome. Predicting risks
57 and outcomes differs from weighing interventions and intervening. Prediction calculates a future
58 event in the absence of any action or change; intervention presumes an enacted choice that may
59 influence the future, which requires consideration of the underlying causal structure.
60 Intervention imagines how the world would be if we made different choices, e.g. *“would the patient be*
61 *cured if we administered amoxicillin instead of a cephalosporin for their upper respiratory tract infection?”* or *“if that*
62 *pre-hypertensive patient had accomplished ten to fifteen minutes of moderate physical activity per day instead of being*
63 *prescribed a diuretic, would they have become hypertensive five years later?”* By asking ourselves what would
64 have been the effect of something if we had not taken an action, or vice versa, we are computing so-
65 called ‘counterfactuals’. Among different counterfactual options, we choose the ones that minimise
66 harm while maximising patient benefit. A similar cognitive process happens when a deep learning
67 machine plays a game and must decide on the next move. Such artificial neural network architecture
68 has been fed the game ruleset, millions of game scenarios, and has learned through trial and error by
69 playing against other human players, networks or even against itself. In each move of the game, the
70 machine chooses the best counterfactual move based on its domain knowledge ^{3,4}.
71 With domain knowledge of the variables involved in a hypothesised cause-effect route and enough
72 data generated at random to cover all possible path configurations, it is possible to deduce causal

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

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

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

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

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

108 *Causal inference and counterfactuals*

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

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

121 the conditional probability $P(E|C)$ of observing E after observing C can be different from the
122 interventional probability $P(E|\text{do}(C))$ of observing E after doing C . In RCTs –when C is
123 randomised– the ‘do’ is guaranteed and unconditioned, while with observational data, it is not.
124 Causal calculus helps resolve interventional from conditional probabilities when a causal structure is
125 assumed¹⁶. **Figure 1** illustrates the difference between observing and doing using biomedical target
126 examples.

127 In a nutshell, the major hurdles to ascertaining causal effects from observational data include: the
128 failure to disambiguate interventional from conditional distributions, to identify all potential sources
129 of bias¹⁷ and to select an appropriate functional form for all variables, i.e. model misspecification^{18–}
130²⁰.

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

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

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

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

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

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

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

175 *An old neural network fiasco and a new possible paradox*

176 In 1997, Cooper et al.²⁷ investigated several machine learning models, including rule-based and
177 neural networks, for predicting mortality of hospital patients admitted with pneumonia. The neural
178 network greatly outperformed logistic regression; however, the authors discouraged using black box
179 models in clinical practice. They showed how the rule-based method learned that ‘IF patient
180 admitted (with pneumonia) has history of asthma THEN patient has lower risk of death from
181 pneumonia’²⁸. This counterintuitive association was also later confirmed using generalised additive
182 models²⁹. The physicians explained that patients admitted with pneumonia and known history of
183 asthma would likely be transferred to intensive care and treated aggressively, thus having higher odds
184 of survival than the general population admitted with pneumonia. The authors recommended to
185 employ interpretable models instead of black boxes, to identify counterintuitive, surprising patterns
186 and remove them. At this point, the model development is no longer automated and requires
187 domain knowledge. Upon reflection, those models inferred without modifications, either
188 interpretable or black box, would have worked well at predicting mortality but they could not have
189 been used to test new interventions to reduce mortality, as the recommended actions would have
190 consequentially led to ‘less care’ for asthmatic patients.

191 More recently, a possible data-driven improvement in the evaluation of fall risk in hospitals was
192 investigated³⁰. Standard practice involves a nurse-led evaluation of patients' history of falls,
193 comorbidities, mental health, gait, ambulatory aids, and intravenous therapy summarised with the
194 Morse scale. To assess the predictive ability of the Morse scale (standard practice), its individual
195 components, and new expert-based predictors (e.g. extended clinical profiles and information on
196 hospital staffing levels), a matched study was performed including patients with and without a fall.
197 Logistic regression and decision trees were used. The additional variables hypothesised by the
198 experts were associated with the outcome and all new models yielded higher discrimination than the
199 Morse scale, but a surprising finding was observed: in all configurations, older patients were at a
200 lower risk of falls. This is contrary to current expert's knowledge, which associates older age with
201 increased frailty and therefore fall risk. If such model were used for intervention, it would not
202 prioritise the elderly for fall prevention –a potentially devastating consequence of data-driven
203 inference. It is uncertain if this old age paradox is due to a bias. One possible explanation is that
204 older patients are usually monitored and aided more frequently because they are indeed at higher
205 risk, while younger people may be more independent and less prone to accept assistance. Other
206 issues at play could be survivorship bias, selectively unreported falls, and study design. One possible
207 approach to bias reduction is to design the study and extract the data by simulating an RCT, where
208 causal effects on “randomised” interventions can be estimated directly, as we discuss in the next
209 section.

210 *The target trial*

211 Target trials refer to RCTs that can be emulated using data from large, observational databases to
212 answer causal questions of comparative treatment effect³¹. Although RCTs are the gold standard for
213 discerning causal effects, there exists many scenarios in which they are neither feasible nor ethical to
214 conduct. Alternatively, observational data appropriately adjusted for measured confounding bias –

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

221 As an example of the target trial framework, data from public surveillance and clinical claims
222 repositories were used to replicate two RCTs, one investigating treatment effects on colorectal
223 cancer and the other on pancreatic adenocarcinoma³². Each study explicitly adhered to the target
224 trial framework, deviating from the RCT design only in the assignment procedures, justifiably due to
225 lack of randomisation. The results were consistent with the target trials—all of which reported a null
226 effect. In contrast, when the authors modelled the treatment effects using a non-RCT-like study
227 design with the same variables, the mortality estimates were both inconsistent with the target trials.
228 These examples demonstrate the need to uphold target trial design in the investigation of treatment
229 effects using observational data. Moreover, coupled with machine learning methods equipped to
230 extrapolate more useful information from big data sources, the target trial framework has the
231 potential to serve as the foundation for exploring causal processes currently unknown.

232 *Causal effect inference and automated learning of causal structures*

233 Prediction models inferred automatically using data without any domain knowledge—from linear
234 regression to deep learning— only approximate a function, and do not necessarily hold a causal
235 meaning. Instead, by estimating interventional in place of conditional probabilities, models can
236 reproduce causal mechanisms. Through counterfactual predictions, models become interventional,
237 actionable, and avoid the pitfalls such as those described in the pneumonia and fall risk examples. In
238 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 matching⁴². Often, these estimators can be coupled with
252 machine learning, e.g. causal decision trees⁴³, Bayesian regression trees⁴⁴, and random forests for
253 estimating individual treatment effects⁴⁵. 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 proposed⁴⁶.
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 theorem⁴⁷– led to the
262 development of automated search and ambiguity resolution of causal structures from data, through

263 so-called patterns and partial ancestral graphs. When assumptions are met (Markov/causal
264 faithfulness⁴⁸), there are asymptotically correct procedures that can predict an effect or raise an
265 ambiguity, and determine graph equivalence⁴⁹. However, the probability of an effect cannot be
266 obtained without deciding on a prior distribution of the graphs and parameters. Also, the number of
267 graphs is super-exponential in the number of observed variables and may even be infinite with
268 hidden variables⁵⁰, making an exhaustive search computationally unfeasible⁵¹. Today, several
269 heuristic methods for causal structure search are available, from the PC algorithm that assumes
270 causal sufficiency, to others like FCI or RFCI that extend to latent variables^{52,53}.

271 The enrichment of deep learning with causal methods also provides interesting new insights to
272 address bias. For instance, a theoretical analysis for identification of individual treatment effect
273 under strong ignorability has been derived⁵⁴, and an approach to exploit instrumental variables for
274 counterfactual prediction within deep learning is also available⁵⁵.

275 *Model transportability and prediction invariance*

276 Validation of causal effects under determined causal structures is especially needed when such
277 effects are estimated in limited settings, e.g. RCTs. Transportability is a data fusion framework for
278 external validation of intervention models and counterfactual queries. As defined by Pearl and
279 Bareinboim⁵⁶, transportability is a “license to transfer causal effects learned in experimental studies
280 to a new population, in which only observational studies can be conducted.” By combining datasets
281 generated under heterogeneous conditions, transportability provides formal mathematical tools to (i)
282 evaluate whether results from one study (e.g. a causal relationship identified in an RCT) could be
283 used to generate valid estimates of the same causal effect in another study of different setting (e.g. an
284 observational study of the same causal effect in a different population); and (ii) estimate what the
285 causal effect would have been if the study had been conducted in the new setting^{57,58}. The
286 framework utilises selection diagrams⁵⁹, encoding the causal relationships of variables of interest in a

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

292 One of Pearl’s transportability examples is shown in **Figure 4**. In this example, a RCT is conducted
 293 in city \mathcal{A} (original environment) and a causal effect of treatment x on outcome y , $P(y|do(x))$, is
 294 determined. We wish to generalise if the treatment works also in the population of city B (target
 295 environment) where only observational data is available, since it happens that the age distribution in
 296 city \mathcal{A} , $P(z)$, is different than that $P^*(z)$ in city B . The city B specific x -to- y causal effect
 297 $P^*(y|do(x))$ is estimated as:

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

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

301 On the other hand, a causal effect is not always transportable. Following the example above, the x -
 302 to- y causal effect is not transportable from City \mathcal{A} to City B if only the overall causal effect
 303 $P(y|do(x))$ is known whereas the age-specific causal effect $P(y|do(x), z)$ is unknown.

304 Transportability theory is being extended to a variety of more complex causal relationships, e.g.
 305 sample selection bias⁵⁸, leaping forward from toy examples to real-world problems⁶⁰. Therefore –
 306 linking back with the problematic examples we discussed in the previous sections– one could use
 307 transportability to determine how the asthma or old age effects are/are not transportable from one
 308 population to another. It is also worth noting how transportability evokes the field of domain

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

322 *Conclusions*

323 We explored common pitfalls of data-driven developments in machine learning for healthcare,
324 distinguishing between prediction and intervention models that are actionable in support of clinical
325 decision-making. Importantly, the development of intervention models requires careful
326 consideration of causality. Hernan *et al.* ⁶⁴ commented that “a recent influx of data analysts, many
327 not formally trained in statistical theory, bring a fresh attitude that does not *a priori* exclude causal
328 questions” yet called –and we strongly endorse such call– for training curricula in data science that
329 well-differentiate descriptive, prediction, and intervention modelling.
330 Undertaking causal machine learning is key to ethical artificial intelligence for healthcare, equivalent
331 to a doctor’s oath to “first do no harm” ⁶⁵. Healthcare intervention models involve actionable inputs
332 and need –implicitly or explicitly– to model causal pathways to compute the correct counterfactuals.

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

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

355

356 **Acknowledgments**

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

362

363 **Author Contributions**

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

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

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

368

369 **Competing Interests Statement**

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

372

373 **References**

- 374 1. Norgeot, B., Glicksberg, B. S. & Butte, A. J. A call for deep-learning healthcare. *Nature*
375 *Medicine* (2019). doi:10.1038/s41591-018-0320-3
- 376 2. Wiens, J. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat.*
377 *Med.* (2019). doi:10.1038/s41591-019-0548-6
- 378 3. Silver, D. *et al.* Mastering the game of Go without human knowledge. *Nature* (2017).
379 doi:10.1038/nature24270
- 380 4. Jin, P., Keutzer, K. & Levine, S. Regret minimization for partially observable deep

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

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

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

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

388 8. Christodoulou, E. *et al.* A systematic review shows no performance benefit of machine
389 learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*
390 (2019). doi:10.1016/j.jclinepi.2019.02.004

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

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

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

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

402 13. Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C. & Faisal, A. A. The Artificial
403 Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med.*
404 (2018). doi:10.1038/s41591-018-0213-5

- 405 14. Lewis, D. K. Causation. *J. Philos.* **70**, (1973).
- 406 15. Mackie, J. L. *The Cement of the Universe*. (Oxford, Clarendon Press, 1974).
- 407 16. Pearl, J. *Causality: Models, Reasoning and Inference*. (Cambridge University Press, 2009).
- 408 17. Rothman, K. J., Greenland, S. & Lash, T. *Modern Epidemiology, 3rd Revised Edition*. Lippincott
409 *Williams & Williams* (2012). doi:10.1017/CBO9781107415324.004
- 410 18. Lehmann, E. L. Model Specification: The views of Fisher and Neyman, and later
411 developments. *Stat. Sci.* (1990). doi:10.1214/ss/1177012164
- 412 19. Vansteelandt, S., Bekaert, M. & Claeskens, G. On model selection and model misspecification
413 in causal inference. *Stat. Methods Med. Res.* (2012). doi:10.1177/0962280210387717
- 414 20. Asteriou, D., Hall, S. G., Asteriou, D. & Hall, S. G. Misspecification: Wrong Regressors,
415 Measurement Errors and Wrong Functional Forms. in *Applied Econometrics* (2016).
416 doi:10.1057/978-1-137-41547-9_8
- 417 21. Sackett, D. L. Bias in analytic research. *J. Chronic Dis.* (1979). doi:10.1016/0021-
418 9681(79)90012-2
- 419 22. Banack, H. R. & Kaufman, J. S. The 'obesity paradox' explained. *Epidemiology* (2013).
420 doi:10.1097/EDE.0b013e31828c776c
- 421 23. Pearl, J. Causal diagrams for empirical research. *Biometrika* (1995).
422 doi:10.1093/biomet/82.4.669
- 423 24. Greenland, S., Pearl, J. & Robins, J. M. Causal diagrams for epidemiologic research.
424 *Epidemiology* (1999). doi:10.1097/00001648-199901000-00008
- 425 25. Westreich, D. & Greenland, S. The table 2 fallacy: Presenting and interpreting confounder
426 and modifier coefficients. *American Journal of Epidemiology* (2013). doi:10.1093/aje/kws412
- 427 26. Wei, L., Brookhart, M. A., Schneeweiss, S., Mi, X. & Setoguchi, S. Implications of m bias in
428 epidemiologic studies: A simulation study. *Am. J. Epidemiol.* (2012). doi:10.1093/aje/kws165

- 429 27. Cooper, G. F. *et al.* An evaluation of machine-learning methods for predicting pneumonia
430 mortality. *Artif. Intell. Med.* (1997). doi:10.1016/S0933-3657(96)00367-3
- 431 28. Ambrosino, R., Buchanan, B. G., Cooper, G. F. & Fine, M. J. The use of misclassification
432 costs to learn rule-based decision support models for cost-effective hospital admission
433 strategies. *Proc. Annu. Symp. Comput. Appl. Med. Care* (1995).
- 434 29. Caruana, R. *et al.* Intelligible models for healthcare: Predicting pneumonia risk and hospital
435 30-day readmission. in *Proceedings of the ACM SIGKDD International Conference on Knowledge
436 Discovery and Data Mining* (2015). doi:10.1145/2783258.2788613
- 437 30. Lucero, R. J. *et al.* A data-driven and practice-based approach to identify risk factors
438 associated with hospital-acquired falls: Applying manual and semi- and fully-automated
439 methods. *Int. J. Med. Inform.* (2019). doi:10.1016/j.ijmedinf.2018.11.006
- 440 31. Hernán, M. A. & Robins, J. M. Using Big Data to Emulate a Target Trial When a
441 Randomized Trial Is Not Available. *Am. J. Epidemiol.* (2016). doi:10.1093/aje/kwv254
- 442 32. Petito, L. C. *et al.* Estimates of Overall Survival in Patients With Cancer Receiving Different
443 Treatment Regimens: Emulating Hypothetical Target Trials in the Surveillance,
444 Epidemiology, and End Results (SEER)–Medicare Linked Database. *JAMA Netw. Open* **3**,
445 e200452–e200452 (2020).
- 446 33. Pearl, J. Causal diagrams for empirical research. *Biometrika* (1995).
447 doi:10.1093/biomet/82.4.669
- 448 34. Westland, J. C. An introduction to structural equation models. in *Studies in Systems, Decision and
449 Control* (2019). doi:10.1007/978-3-030-12508-0_1
- 450 35. Bollen, K. A. & Pearl, J. Eight Myths About Causality and Structural Equation Models. in
451 (2013). doi:10.1007/978-94-007-6094-3_15
- 452 36. Hernán, M. A. & Robins, J. M. Estimating causal effects from epidemiological data. *Journal of*

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

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

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

458 39. Rose, S. & van der Laan, M. J. Targeted Learning: Causal Inference for Observational and
459 Experimental Data. *Target. Learn. Causal Inference Obs. Exp. Data* (2011). doi:10.1007/978-1-
460 4419-9782-1

461 40. Naimi, A. I., Cole, S. R. & Kennedy, E. H. An introduction to g methods. *Int. J. Epidemiol.*
462 (2017). doi:10.1093/ije/dyw323

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

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

467 43. Li, J., Ma, S., Le, T., Liu, L. & Liu, J. Causal Decision Trees. *IEEE Trans. Knowl. Data Eng.*
468 (2017). doi:10.1109/TKDE.2016.2619350

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

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

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

- 477 47. Verma, T. & Pearl, J. Causal Networks: Semantics and Expressiveness. in *Machine Intelligence*
478 *and Pattern Recognition* (1990). doi:10.1016/B978-0-444-88650-7.50011-1
- 479 48. Jaber, A., Zhang, J. & Bareinboim, E. Causal identification under Markov equivalence. in *34th*
480 *Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018* (2018).
- 481 49. Richardson, T. Equivalence in Non-Recursive Structural Equation Models. in *Compstat* (1994).
482 doi:10.1007/978-3-642-52463-9_59
- 483 50. Heckerman, D., Meek, C. & Cooper, G. F. A Bayesian approach to causal discovery. *Studies in*
484 *Fuzziness and Soft Computing* (2006). doi:10.1007/10985687_1
- 485 51. Peter Spirtes, C. G. and R. S. Causation, Prediction, and Search. 2nd edn. MIT Press. *Stat.*
486 *Med.* (2003). doi:10.1002/sim.1415
- 487 52. Glymour, C., Zhang, K. & Spirtes, P. Review of causal discovery methods based on graphical
488 models. *Front. Genet.* (2019). doi:10.3389/fgene.2019.00524
- 489 53. Colombo, D. & Maathuis, M. H. Order-independent constraint-based causal structure
490 learning. *J. Mach. Learn. Res.* (2015).
- 491 54. Shalit, U., Johansson, F. D. & Sontag, D. Estimating individual treatment effect:
492 Generalization bounds and algorithms. in *34th International Conference on Machine Learning,*
493 *ICML 2017* (2017).
- 494 55. Hartford, J., Lewis, G., Leyton-Brown, K. & Taddy, M. Deep {IV}: A Flexible Approach for
495 Counterfactual Prediction. in *Proceedings of the 34th International Conference on Machine Learning*
496 (eds. Precup, D. & Teh, Y. W.) **70**, 1414–1423 (PMLR, 2017).
- 497 56. Pearl, J. & Bareinboim, E. External validity: From do-calculus to transportability across
498 populations. *Stat. Sci.* (2014). doi:10.1214/14-STS486
- 499 57. Dahabreh, I. J., Robertson, S. E., Tchetgen, E. J., Stuart, E. A. & Hernán, M. A. Generalizing
500 causal inferences from individuals in randomized trials to all trial-eligible individuals.

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

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

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

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

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

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

515 63. Subbaswamy, A., Schulam, P. & Saria, S. Preventing Failures Due to Dataset Shift: Learning
516 Predictive Models That Transport. BT - The 22nd International Conference on Artificial
517 Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan. 3118–
518 3127 (2019).

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

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

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

- 525 67. Kusner, M. J. & Loftus, J. R. The long road to fairer algorithms. *Nature* (2020).
526 doi:10.1038/d41586-020-00274-3
- 527 68. van Amsterdam, W. A. C., Verhoeff, J. J. C., de Jong, P. A., Leiner, T. & Eijkemans, M. J. C.
528 Eliminating biasing signals in lung cancer images for prognosis predictions with deep
529 learning. *npj Digit. Med.* (2019). doi:10.1038/s41746-019-0194-x
- 530 69. Moons, K. G. M. *et al.* Transparent reporting of a multivariable prediction model for
531 individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann. Intern. Med.*
532 (2015). doi:10.7326/M14-0698

533

534

535 **Figure Legends**

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

546 **Figure 2. *Examples of confounding bias and collider bias.*** Confounding (panel a) can occur
547 when there exists a common cause for both exposure and outcome, while a collider (panel b)
548 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.

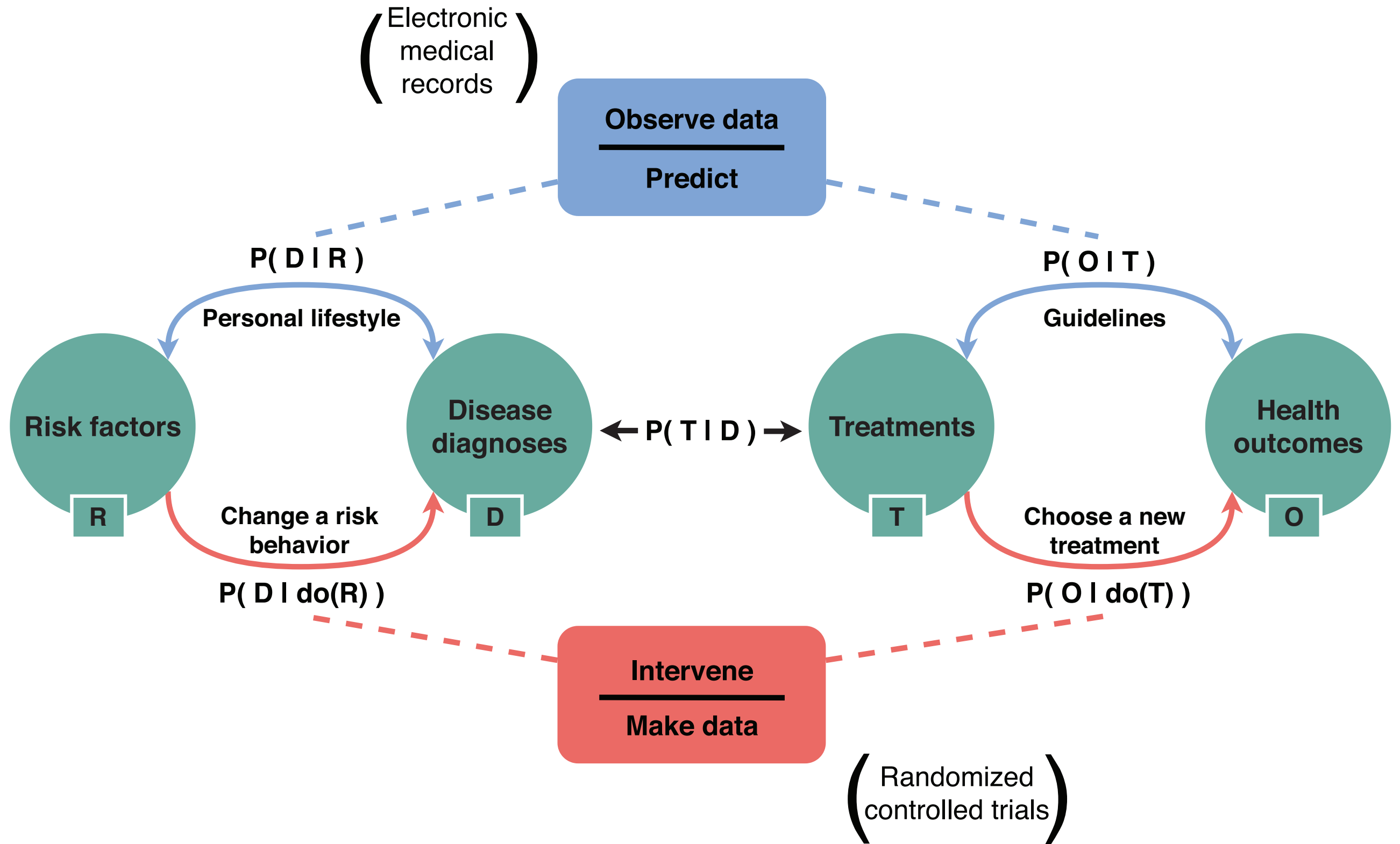
558

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



a

Model inputs

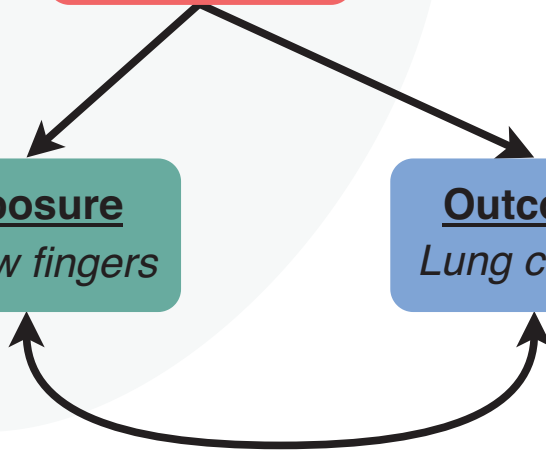
Confounder
Smoking

Exposure

Yellow fingers

Outcome

Lung cancer



Biased association when
confounder is not included

b

Model inputs

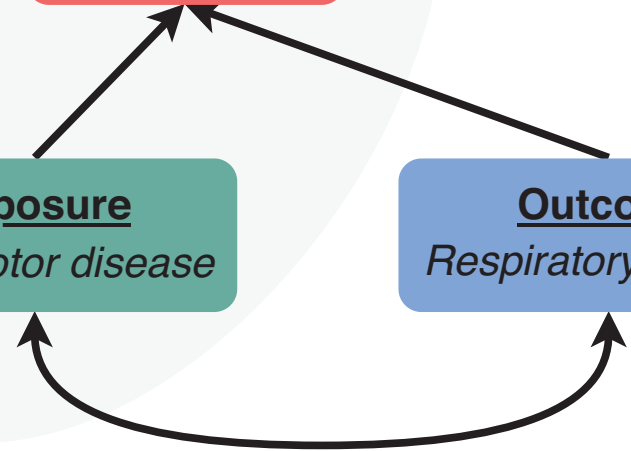
Collider
Hospitalization

Exposure

Locomotor disease

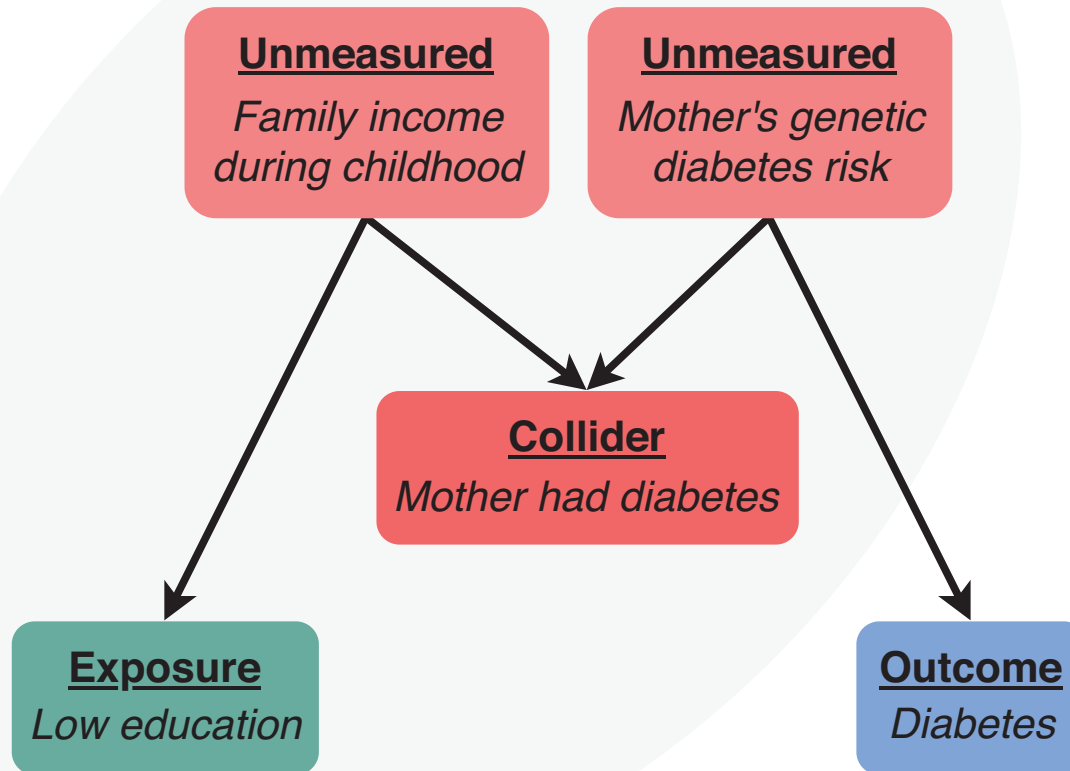
Outcome

Respiratory disease



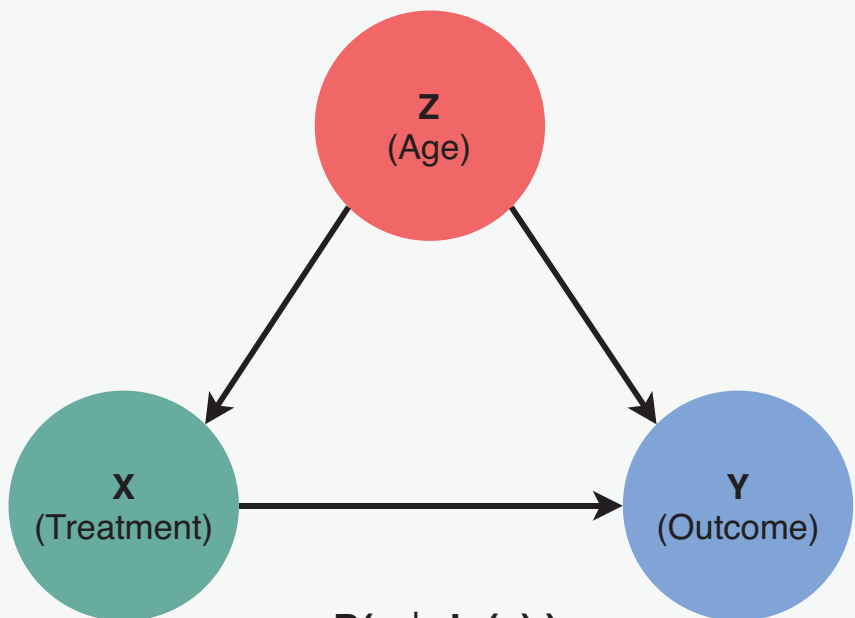
Biased association when
collider is included

Model inputs



a

Original environment
(City A, interventional)



$P(y | do(x))$

* $P(y | do(x), z)$ is known



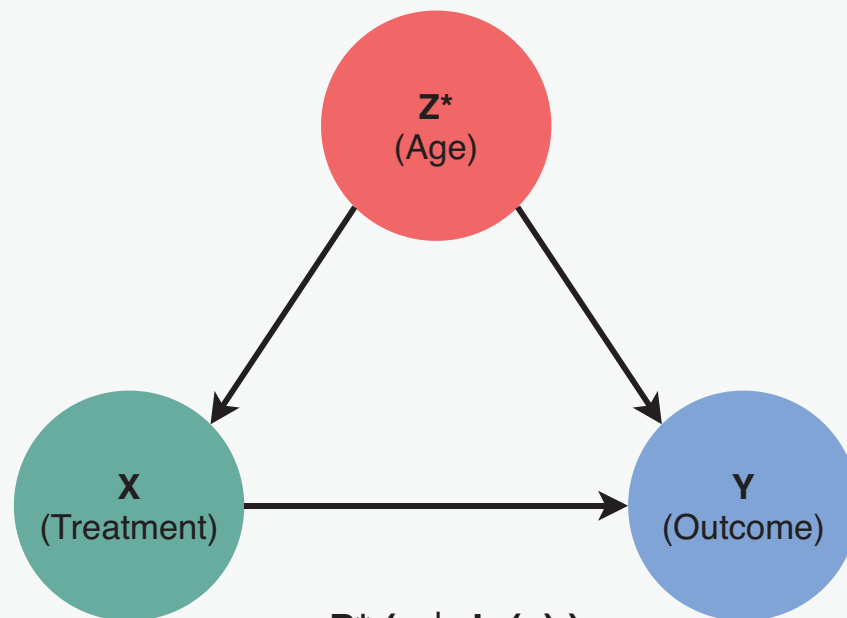
Transport



* $P(y | do(x), z)$ is unknown

b

Target environment
(City B, observational)



$P^*(y | do(x))$