The past, present and future of causal learning

Hedong Yan

CS, HKBU

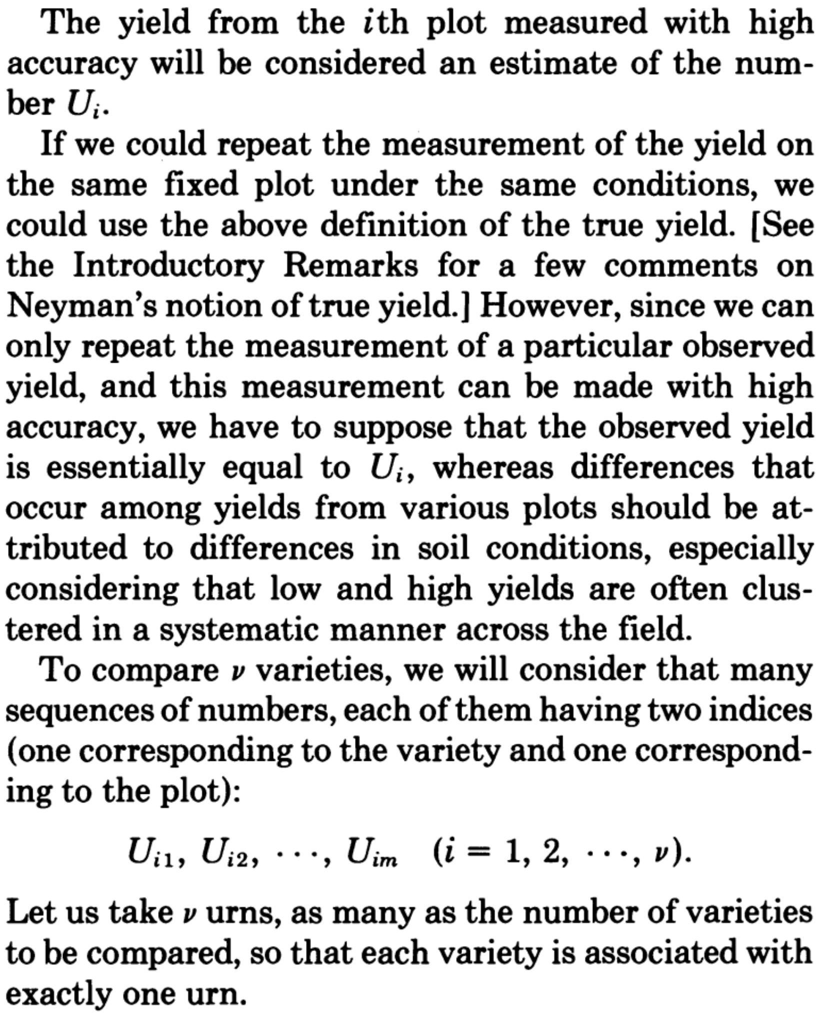
Roadmap

- Past
  - Neyman
  - Potential Outcome
- Present
  - Evaluation
- Future

History of causal inference (Neyman)

- Jerzy Neyman was a ***Polish*** mathematician and statistician who spent the first part of his professional career at various institutions in Warsaw, ***Poland*** and then at ***University College London***, and the second part at the ***University of California, Berkeley***.
- He published many books dealing with experiments and statistics, and devised ***the way which the FDA tests medicines*** today.

History of causal inference (Neyman)



The yield from the ith plot measured with high accuracy will be considered an estimate of the number U_i.

If we could repeat the measurement of the yield on the same fixed plot under the same conditions, we could use the above definition of the true yield. [See the Introductory Remarks for a few comments on Neyman's notion of true yield.] However, since we can only repeat the measurement of a particular observed yield, and this measurement can be made with high accuracy, we have to suppose that the observed yield is essentially equal to U_i, whereas differences that occur among yields from various plots should be attributed to differences in soil conditions, especially considering that low and high yields are often clustered in a systematic manner across the field.

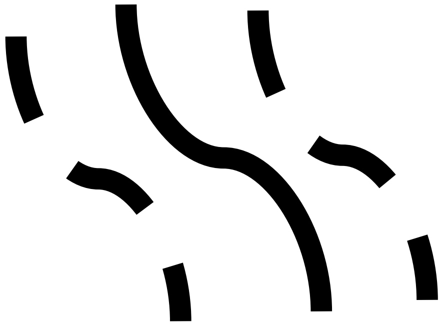
To compare ν varieties, we will consider that many sequences of numbers, each of them having two indices (one corresponding to the variety and one corresponding to the plot):

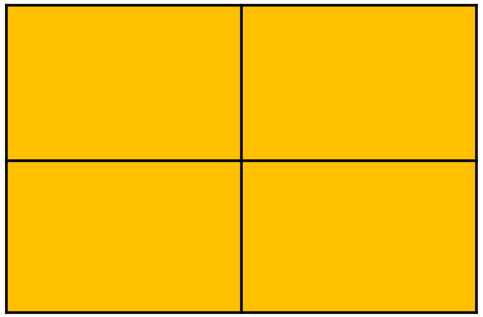
$$U_{i1}, U_{i2}, \cdots, U_{im} \quad (i = 1, 2, \cdots, \nu).$$

Let us take ν urns, as many as the number of varieties to be compared, so that each variety is associated with exactly one urn.

- Unknown potential yields

Plots





Crop Varieties

Splawa-Neyman, Jerzy, Dorota M. Dabrowska, and Terrence P. Speed. "On the application of probability theory to agricultural experiments. Essay on principles. Section 9." *Statistical Science* (1990): 465-472.

History of causal inference (Neyman)

- “We see that knowledge of the preceding trials has an effect on the probability of outcomes of subsequent trials so that trials conducted in this way are not ***independent***.”
- **Infinite plots:** “Therefore the trials will turn out to be ***independent***, and we will be able to apply the law of large numbers, and our definition of a ***true yield***, and along with it known formulas from probability theory.”

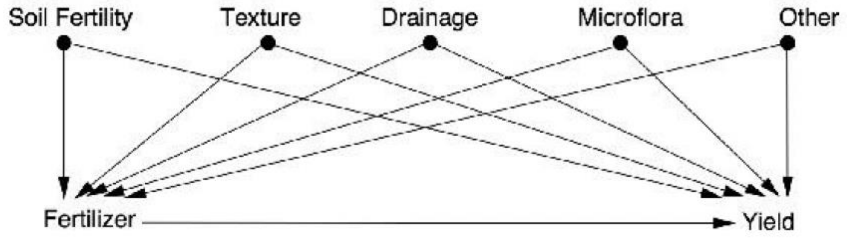


FIGURE 4.4. Model 1: an improperly controlled experiment.

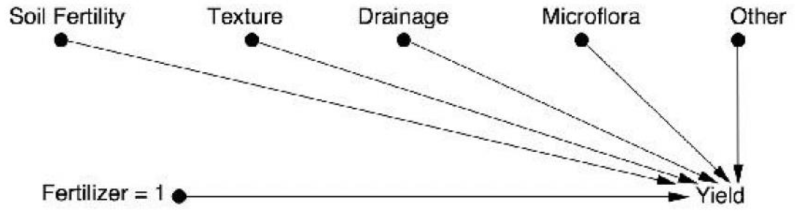


FIGURE 4.5. Model 2: the world we would like to know about.

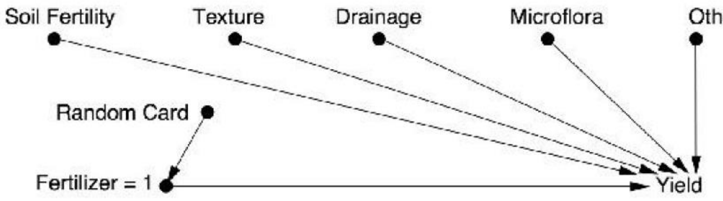


FIGURE 4.6. Model 3: the world simulated by a randomized controlled trial.

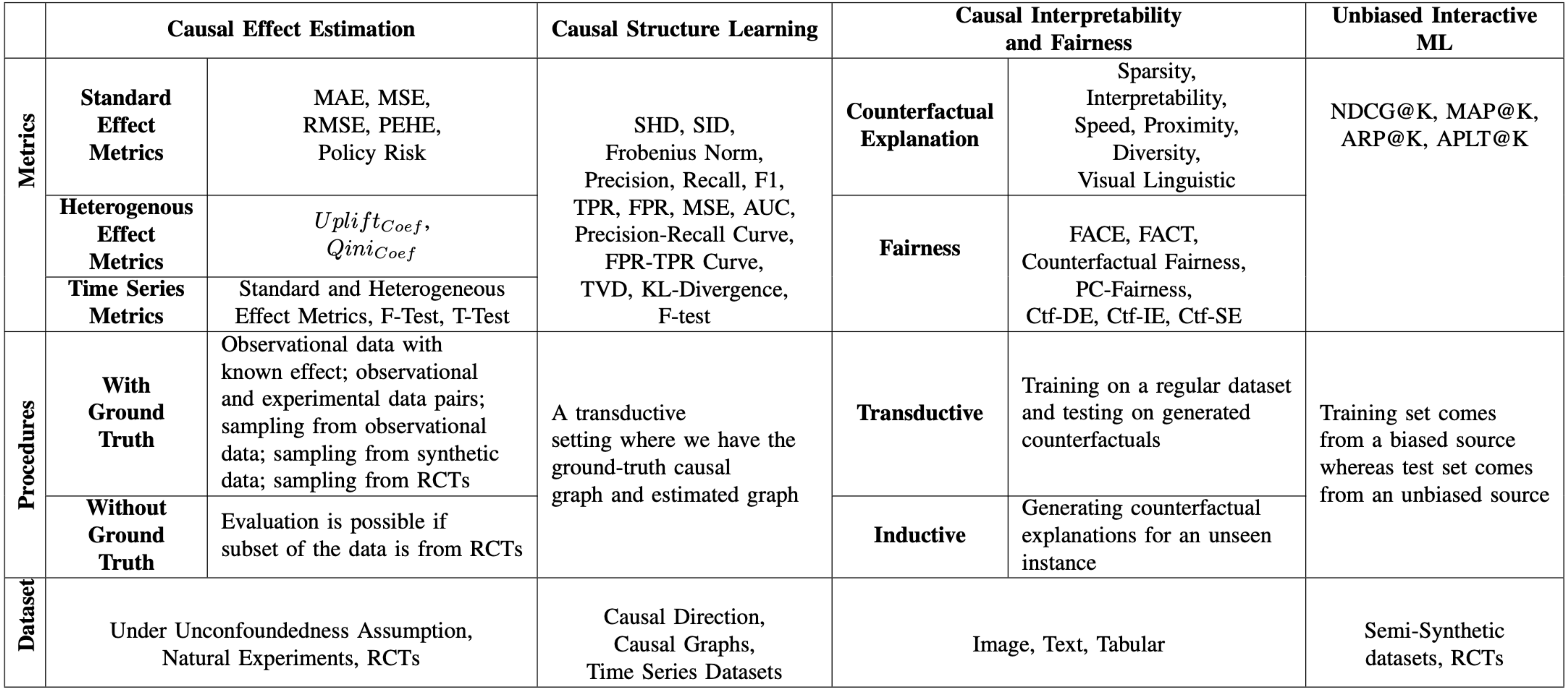
Splawa-Neyman, Jerzy, Dorota M. Dabrowska, and Terrence P. Speed. "On the application of probability theory to agricultural experiments. Essay on principles. Section 9." *Statistical Science* (1990): 465-472.
Pearl, Judea, and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

History of causal inference (Potential Outcome)

- Fundamental problem: ***missing outcomes >= 50%***
- “The first step toward addressing observational studies is to relax the classical randomized experiment assumption that the probability of treatment assignment is a known function. We do maintain, however, in this part of the text, the ***unconfoundedness*** assumption that states that assignment is free from dependence on the potential outcomes. Moreover, we continue to assume that the assignment mechanism is ***individualistic***, so that the probability for unit i is essentially a function of the pre-treatment variables for unit i only, free of dependence on the values of pre-treatment variables for other units. We also maintain the assumption that the assignment mechanism is ***probabilistic***, so that the probability of receiving any level of the treatment is strictly between zero and one for all units.”

Imbens, Guido W., and Donald B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Present of causal learning (Evaluation)



		Causal Effect Estimation		Causal Structure Learning	Causal Interpretability and Fairness		Unbiased Interactive ML
Metrics	**Standard Effect Metrics**	MAE, MSE, RMSE, PEHE, Policy Risk		SHD, SID, Frobenius Norm, Precision, Recall, F1, TPR, FPR, MSE, AUC, Precision-Recall Curve, FPR-TPR Curve, TVD, KL-Divergence, F-test	**Counterfactual Explanation**	Sparsity, Interpretability, Speed, Proximity, Diversity, Visual Linguistic	NDCG@K, MAP@K, ARP@K, APLT@K
	Heterogenous Effect Metrics	$Uplift_{Coef}$, $Qini_{Coef}$			**Fairness**	FACE, FACT, Counterfactual Fairness, PC-Fairness, Ctf-DE, Ctf-IE, Ctf-SE	
	Time Series Metrics	Standard and Heterogeneous Effect Metrics, F-Test, T-Test					
Procedures	**With Ground Truth**	Observational data with known effect; observational and experimental data pairs; sampling from observational data; sampling from synthetic data; sampling from RCTs		A transductive setting where we have the ground-truth causal graph and estimated graph	**Transductive**	Training on a regular dataset and testing on generated counterfactuals	Training set comes from a biased source whereas test set comes from an unbiased source
	Without Ground Truth	Evaluation is possible if subset of the data is from RCTs			**Inductive**	Generating counterfactual explanations for an unseen instance	
Dataset	Under Unconfoundedness Assumption, Natural Experiments, RCTs			Causal Direction, Causal Graphs, Time Series Datasets	Image, Text, Tabular		Semi-Synthetic datasets, RCTs

TABLE I: Summary of metrics, procedures, datasets for evaluating CL approaches.

Cheng, Lu, et al. "Evaluation methods and measures for causal learning algorithms." *IEEE Transactions on Artificial Intelligence* 3.6 (2022): 924-943.

Evaluation Procedures (Evaluation)

- Observation data with known effect (low data availability)
  - causal direction
- Pair of observations and experiments (low data availability)
  - Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens
- Sampling from synthetic data
  - generate observational data from synthetic causal system
  - cannot generalize well to real-world setting
- Sampling from observation data
  - Use known functions to create treatment assignments and outcomes
  - cannot generalize well to real-world setting
- Sampling from RCT
  - OSRCT: treatment assignment is synthetic, middle data availability

Evaluation Procedures (Evaluation)

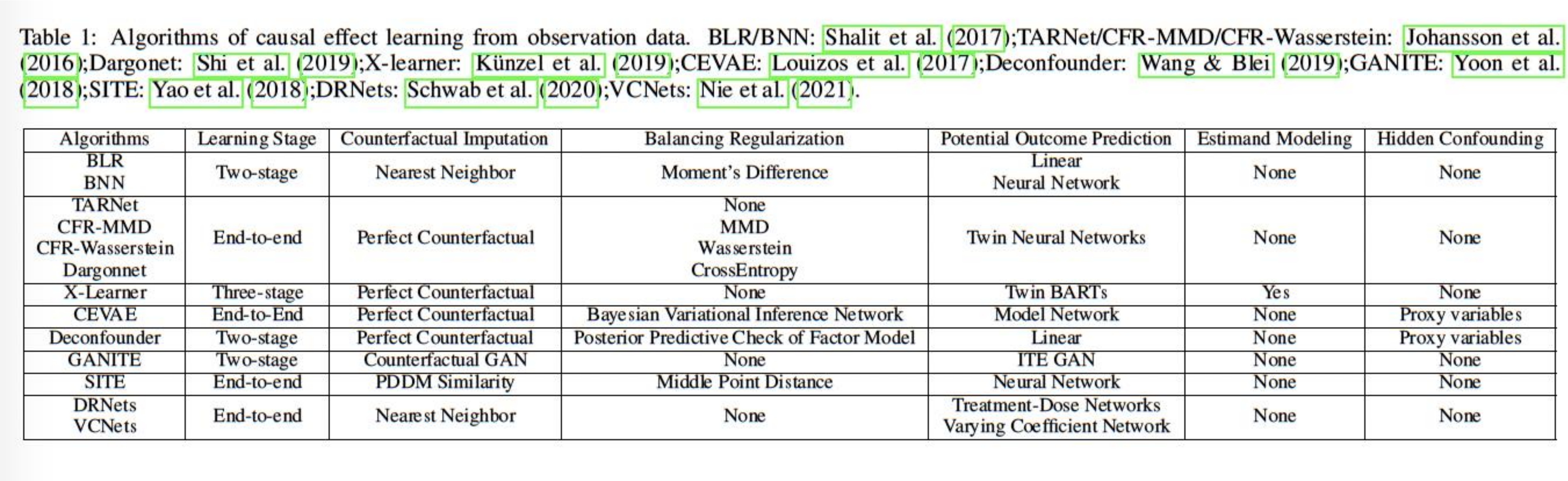
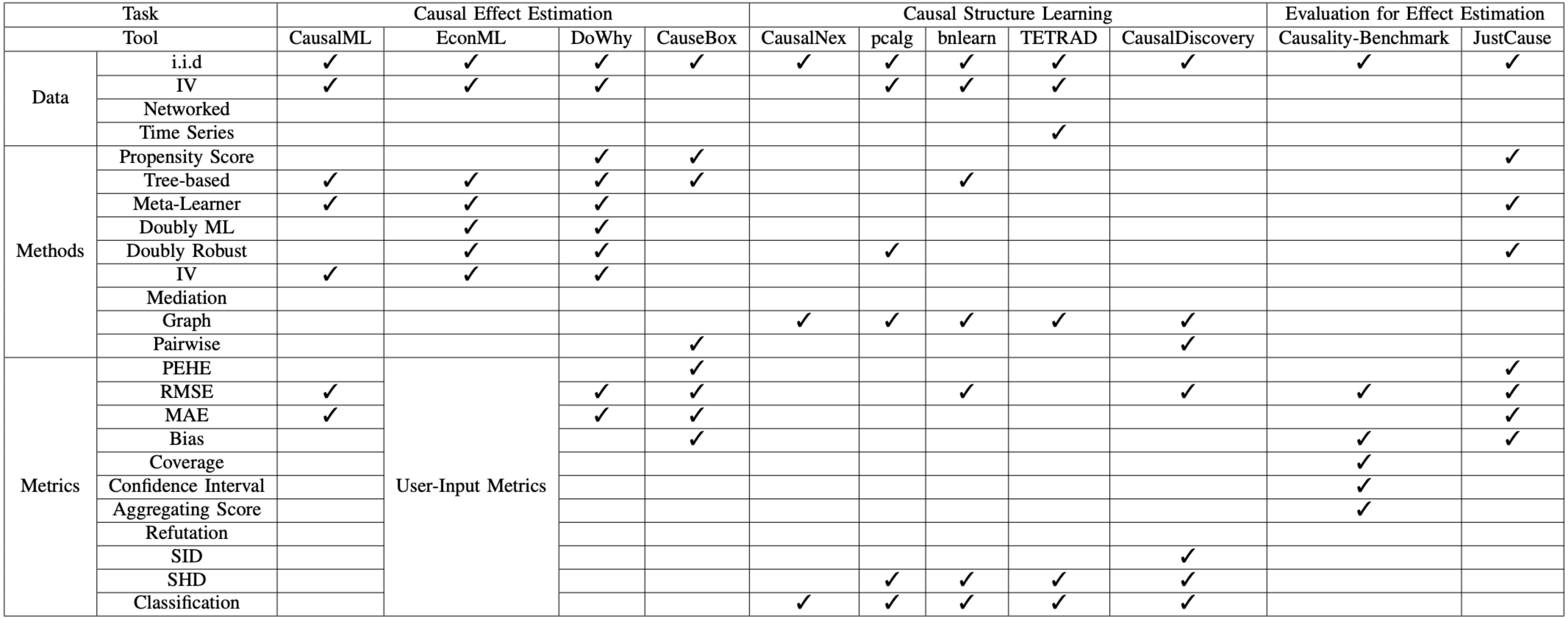


Table 1: Algorithms of causal effect learning from observation data. BLR/BNN: Shalit et al. (2017);TARNet/CFR-MMD/CFR-Wasserstein: Johansson et al. (2016);Dargonet: Shi et al. (2019);X-learner: Künzel et al. (2019);CEVAE: Louizos et al. (2017);Deconfounder: Wang & Blei (2019);GANITE: Yoon et al. (2018);SITE: Yao et al. (2018);DRNets: Schwab et al. (2020);VCNets: Nie et al. (2021).

Algorithms	Learning Stage	Counterfactual Imputation	Balancing Regularization	Potential Outcome Prediction	Estimand Modeling	Hidden Confounding
BLR BNN	Two-stage	Nearest Neighbor	Moment's Difference	Linear Neural Network	None	None
TARNet CFR-MMD CFR-Wasserstein Dargonnet	End-to-end	Perfect Counterfactual	None MMD Wasserstein CrossEntropy	Twin Neural Networks	None	None
X-Learner	Three-stage	Perfect Counterfactual	None	Twin BARTs	Yes	None
CEVAE	End-to-End	Perfect Counterfactual	Bayesian Variational Inference Network	Model Network	None	Proxy variables
Deconfounder	Two-stage	Perfect Counterfactual	Posterior Predictive Check of Factor Model	Linear	None	Proxy variables
GANITE	Two-stage	Counterfactual GAN	None	ITE GAN	None	None
SITE	End-to-end	PDDM Similarity	Middle Point Distance	Neural Network	None	None
DRNets VCNets	End-to-end	Nearest Neighbor	None	Treatment-Dose Networks Varying Coefficient Network	None	None

一定要关注所谓“因果”信息是何时在哪里引入的，来自何处

Present of causal learning (Evaluation)



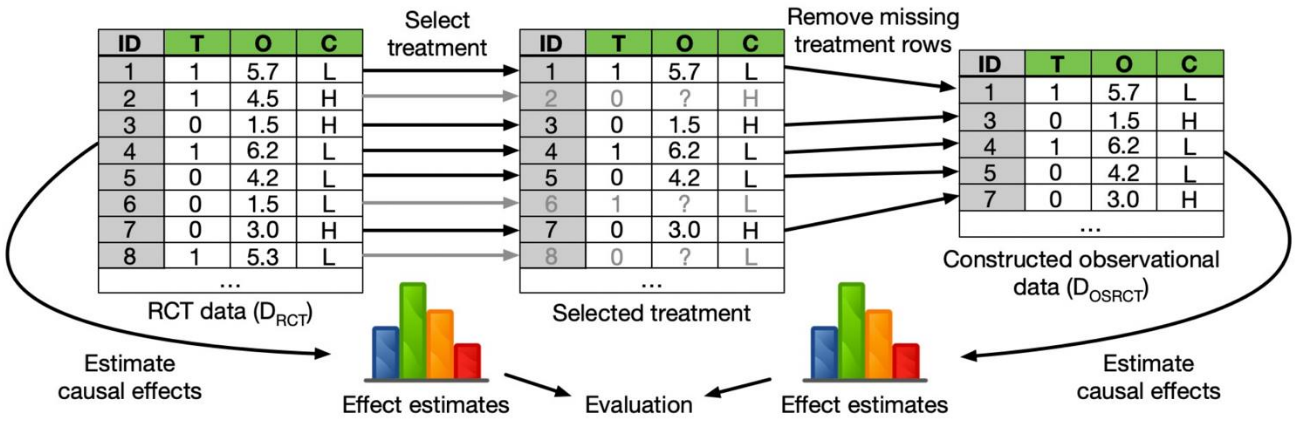
Task		Causal Effect Estimation				Causal Structure Learning					Evaluation for Effect Estimation	
Tool		CausalML	EconML	DoWhy	CauseBox	CausalNex	pcalg	bnlearn	TETRAD	CausalDiscovery	Causality-Benchmark	JustCause
Data	i.i.d	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	IV	✓	✓	✓			✓	✓	✓			
	Networked											
	Time Series								✓			
Methods	Propensity Score			✓	✓							✓
	Tree-based	✓	✓	✓	✓			✓				
	Meta-Learner	✓	✓	✓								✓
	Doubly ML		✓	✓								
	Doubly Robust		✓	✓			✓					✓
	IV	✓	✓	✓								
	Mediation											
	Graph					✓	✓	✓	✓	✓		
	Pairwise				✓					✓		
Metrics	PEHE		User-Input Metrics		✓							✓
	RMSE	✓		✓	✓			✓		✓	✓	✓
	MAE	✓		✓	✓							✓
	Bias				✓						✓	✓
	Coverage										✓	
	Confidence Interval										✓	
	Aggregating Score										✓	
	Refutation											
	SID									✓		
	SHD						✓	✓	✓	✓		
	Classification					✓	✓	✓	✓	✓		

TABLE III: Comparisons of causal inference tools with a focus on the included datasets, methods, and metrics.

Cheng, Lu, et al. "Evaluation methods and measures for causal learning algorithms." *IEEE Transactions on Artificial Intelligence* 3.6 (2022): 924-943.

Present of causal learning (Evaluation)

Data type	Data Availability	Internal Validity	External Validity
All Potential Outcome Data	Usually No	\	\
Randomized Trial Data	Middle	High	Middle
Natural Experiment	Middle	Middle	Middle
Observation Data	Very High	Too many untestable and unrealistic assumptions	\



ID	T	O	C
1	1	5.7	L
2	1	4.5	H
3	0	1.5	H
4	1	6.2	L
5	0	4.2	L
6	0	1.5	L
7	0	3.0	H
8	1	5.3	L

...

RCT data (D_{RCT})

ID	T	O	C
1	1	5.7	L
2	0	?	H
3	0	1.5	H
4	1	6.2	L
5	0	4.2	L
6	1	?	L
7	0	3.0	H
8	0	?	L

...

Selected treatment

ID	T	O	C
1	1	5.7	L
3	0	1.5	H
4	1	6.2	L
5	0	4.2	L
7	0	3.0	H

...

Constructed observational data (D_{OSRCT})

Figure 2. The process of creating observational-style data from a randomized controlled trial.

Gentzel, Amanda M., Purva Pruthi, and David Jensen. "How and why to use experimental data to evaluate methods for observational causal inference." *International Conference on Machine Learning*. PMLR, 2021. https://icml.cc/media/icml-2021/Slides/9159.pdf

Present of causal learning (Evaluation)

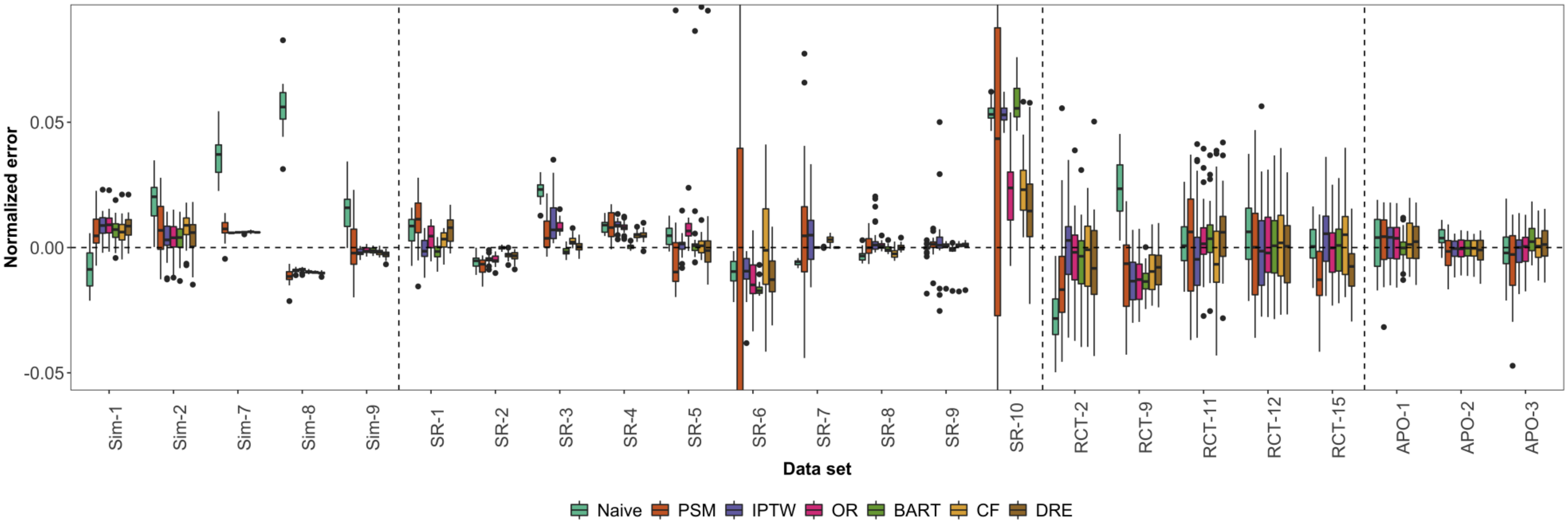


Figure 3. Normalized error in estimating ATE for data sets with continuous outcome. Sim denotes simulator, SR denotes synthetic-response data sets, RCT denotes randomized controlled trials and APO denotes computational systems.

Gentzel, Amanda M., Purva Pruthi, and David Jensen. "How and why to use experimental data to evaluate methods for observational causal inference." *International Conference on Machine Learning*. PMLR, 2021. https://icml.cc/media/icml-2021/Slides/9159.pdf

Present of causal learning (Evaluation)

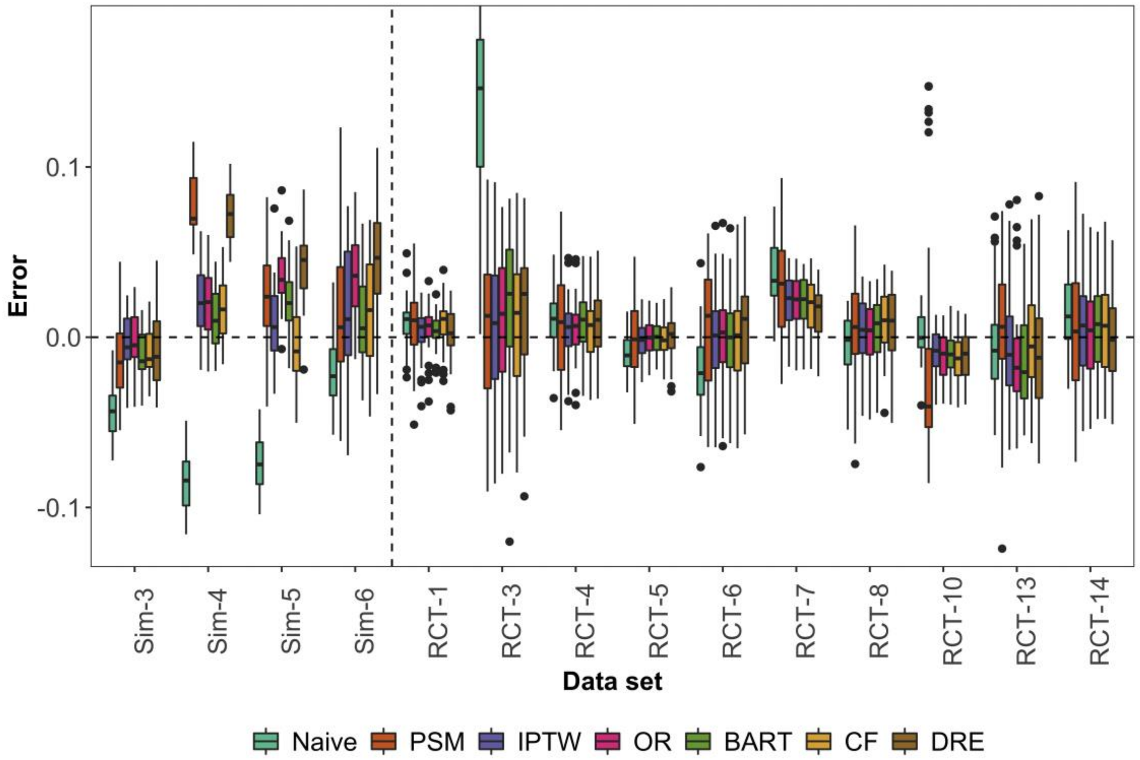


Figure 6. Error in estimating risk difference with two biasing covariates, for data sets with binary outcome

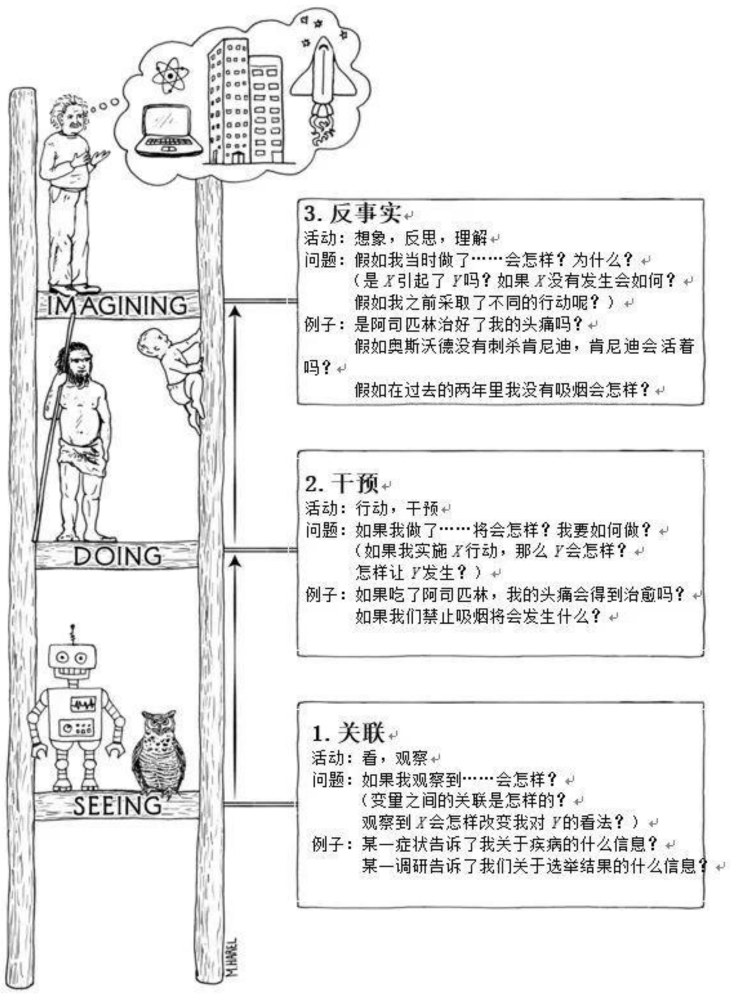
Gentzel, Amanda M., Purva Pruthi, and David Jensen. "How and why to use experimental data to evaluate methods for observational causal inference." *International Conference on Machine Learning*. PMLR, 2021. https://icml.cc/media/icml-2021/Slides/9159.pdf

Now, it is brainstorming time.

Workshop time

Questions:

What is **not** the causality we want?
Why randomized trial is **not** regarded as **not** useful?



3. 反事实
活动：想象，反思，理解
问题：假如我当时做了……会怎样？为什么？
（是 *X* 引起了 *Y* 吗？如果 *X* 没有发生会如何？
假如我之前采取了不同的行动呢？）
例子：是阿司匹林治好了我的头痛吗？
假如奥斯沃德没有刺杀肯尼迪，肯尼迪会活着吗？
假如在过去的两年里我没有吸烟会怎样？

2. 干预
活动：行动，干预
问题：如果我做了……将会怎样？我要如何做？
（如果我实施 *X* 行动，那么 *Y* 会怎样？
怎样让 *Y* 发生？）
例子：如果吃了阿司匹林，我的头痛会得到治愈吗？
如果我们禁止吸烟将会发生什么？

1. 关联
活动：看，观察
问题：如果我观察到……会怎样？
（变量之间的关联是怎样的？
观察到 *X* 会怎样改变我对 *Y* 的看法？）
例子：某一症状告诉了我关于疾病的什么信息？
某一调研告诉了我们关于选举结果的什么信息？

请阅读以下材料简单了解
science和research

https://en.wikipedia.org/wiki/Science
https://en.wikipedia.org/wiki/Research

Principle 1: 从无知出发并用否定之否定获得结论
Principle 2: 追求有用的模型而非正确和真实的模型

Artificial Randomness

"As I know only one thing–that I know nothing" (ignorance)

What is "nothing" that we know for causality?

Randomness is a way to create ignorance. So, we can attain knowledge from it.

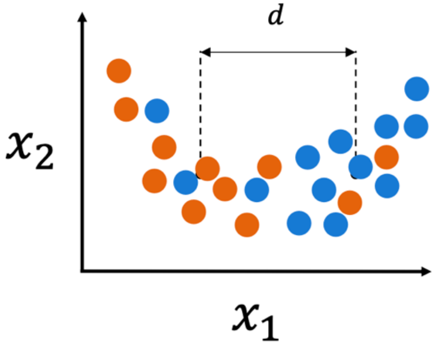
Definition: an assignment that can not be predicted by any model in the list given all pre-treatments of the individual.

背后想法是使用人造随机来代替自然变化莫测的随机性

Use difference method to understand causality

Observational studies:
test ≠ train

- Predict outcome under **unobserved** treatment
- Treatment is **not** assigned equally at random: $p(T=1 \mid X) \neq P(T=1)$
- There is a non-negligible difference between treatment group distributions



Example:
A difference in means

“Treated tend to be younger”

RCT:
Test = Train

想要回答的问题是：如果独立决定是否吃药会怎么样，使用的数据也是独立决定是否吃药时的数据。这里的独立的意思是和所有吃药前的一切可测或不可测变量，可测中的一切已测和未测变量都独立。

如果不考虑总体的不同，对于完美随机试验（全依从，无测量误差，无观察效应，无数据丢失，无噪声），只有治疗分配是不同的。回想Neyman的思想，在随机试验中我们对以下命题有信心“**治疗分配统计独立于任何治疗前变量，无论其是否可测，是否已测**”。
信心来自于哪里？赌场让人相信骰子和硬币的统计独立性，扭蛋抽卡游戏让人相信电脑随机数的统计独立性，物理学让人相信量子方面的统计独立性

因果表示是可能的吗？

因果表示：改变我们的表示，表示的数据值所指向的物理世界中的属性会发生变化，我们期待的物理世界中的受表示影响的输出也会发生变化

这种因果表示是不可能的。

- ✓时间上，**先有的图像，后有的特征**，而非根据随机化的特征采集图像
- ✓图像的改变与物理世界的改变无关
- ✓Fake it till you make it只能让其数值逼近，这种思路最重要的是**把自己想要的这些先于图像产生的东西从自己脑子里提取出来列个清单**

总结

观察数据的价值

- 无论是存在模型对于**治疗分配**预测的特别准，或者是不存在模型/人对于治疗分配预测有帮助都是好事
  - **前者是知道了治**疗分配的机制，后者是知道了治疗分配是随机的
- 观测数据的**治疗前变量的分布是有**意义的，但是观测数据的 Outcome Function是混淆无因果意义的
- 使用**合成的Outcome Function进行评测价值很低**

随机数据的价值

- 无论在什么数据上训练和验证，**任何宣称因果的模型算法的测试必须在随机试验**的衍生数据上
- **采集代价允**许范围内，**随机**试验时采集个体更高维、更多源、更异构的特征，弥补样本量的缺乏，在**保护隐私基础上推动个体随机试验数据公开化**
- RCT不仅应该有统计检验，还应该建立模型预测个体的Outcome，**因果预测测试**中，Confidence > 0.95时，给出**因果模型的认证**，并分发给应用人员根据所属群体的变化自行调整使用
- 模型学习时要重视随机试验中**每个个体单独的Outcome Function**提高因果信息利用率

在做因果问题时，需要时刻问自己随机性来自于哪里，这是它的本质。不可以被定义概率时的sigma代数/可测代数/可测函数的定义所蒙蔽，默认所有随机性来自于分布自身，只满足于可测，拒绝追问随机性的来源。

建议

- 1、把随机试验的 outcome functions 迁移到观测数据，认为存在outcome functions model具有不变性。所有随机性来自于试验对象的covariate，和人造随机数产生的treatment，从而进一步研究covariate的分布变化规律，变成covariate shift问题

- 2、只有观测数据怎么办：学习treatment assignment函数。用所有治疗前可测到的变量无论是哪个个体的去尽可能好地预测treatment，获得一个最好的预测模型。评估这个预测模型的准确度，以二值治疗和f1 score为例，f1 score > 0.95意味着我们很了解treatment assignment，定义一个减法，用学到的个体预测模型减掉治疗变量预测模型，从而得到独立决策时的结果。如果做不到，那么需要想清楚自己想要的是因果关系还是其他东西。如果确定是因果关系，那就采集更多有利于预测的特征和数据，把模型加大去训练，或者做随机试验和找自然实验

欢迎进一步深入的讨论批评!

History of causal inference (Diagram)

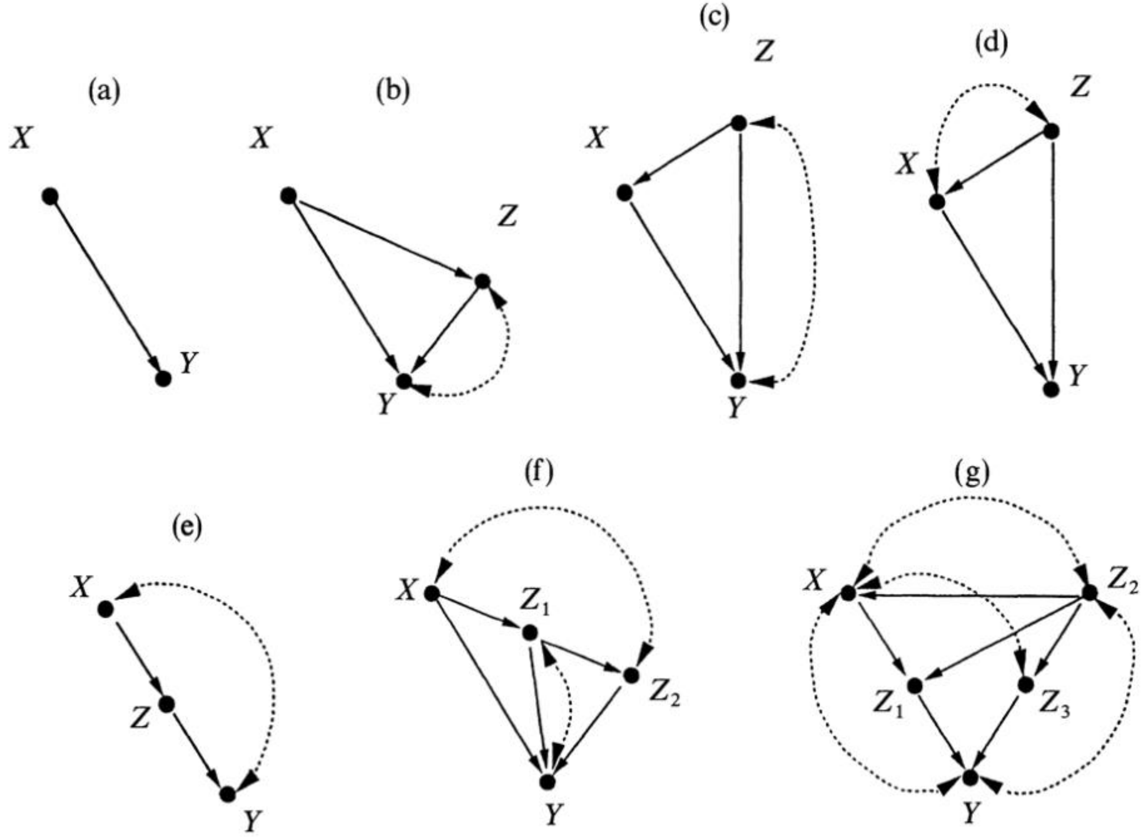
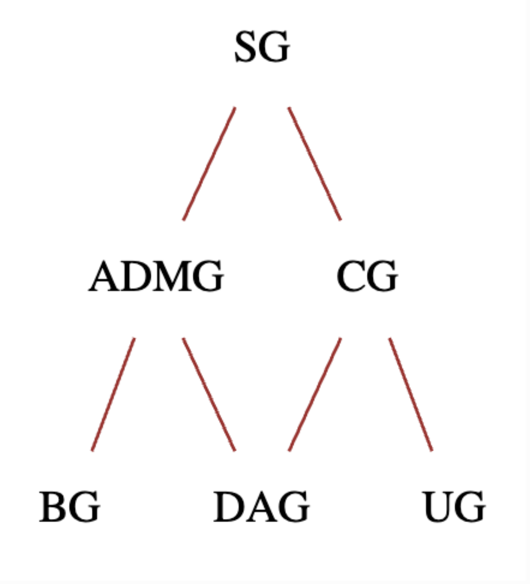


Fig. 6. Typical models in which the effect of X on Y is identifiable. Dashed arcs represent confounding paths, and Z represents observed covariates.

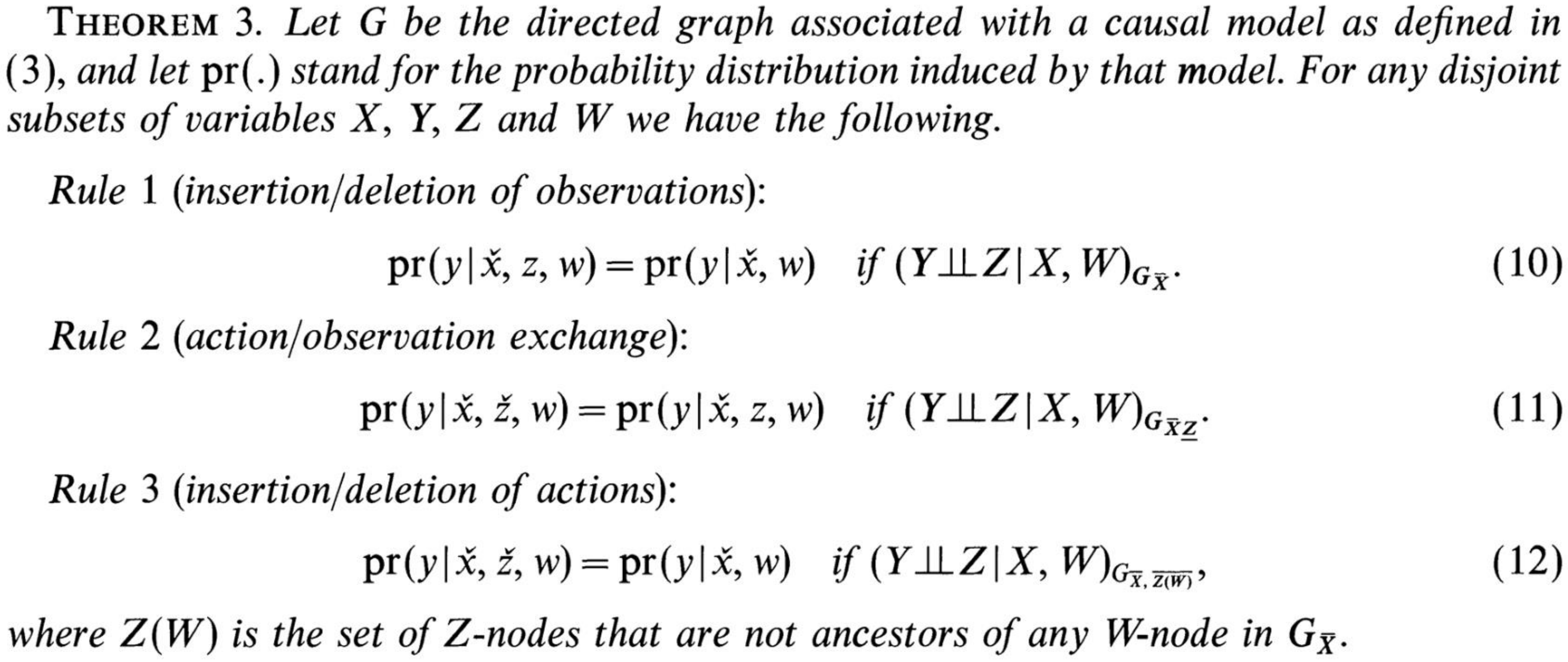
- **Graph**: BG, UG, DAG, ADMG, CG, SG ……
- Do-calculus for computation acceleration



有非常强的启发价值和可视化价值，但批判价值较低

Pearl, Judea. "Causal diagrams for empirical research." Biometrika 82.4 (1995): 669-688.
https://ananke.readthedocs.io/en/latest/notebooks/causal_graphs.html

History of causal inference (Diagram)



THEOREM 3. *Let G be the directed graph associated with a causal model as defined in* (3), *and let* pr(.) *stand for the probability distribution induced by that model. For any disjoint subsets of variables X, Y, Z and W we have the following.*

Rule 1 (*insertion/deletion of observations*):

$$\mathrm{pr}(y \mid \check{x}, z, w) = \mathrm{pr}(y \mid \check{x}, w) \quad \textit{if } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}}}. \tag{10}$$

Rule 2 (*action/observation exchange*):

$$\mathrm{pr}(y \mid \check{x}, \check{z}, w) = \mathrm{pr}(y \mid \check{x}, z, w) \quad \textit{if } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\underline{Z}}}. \tag{11}$$

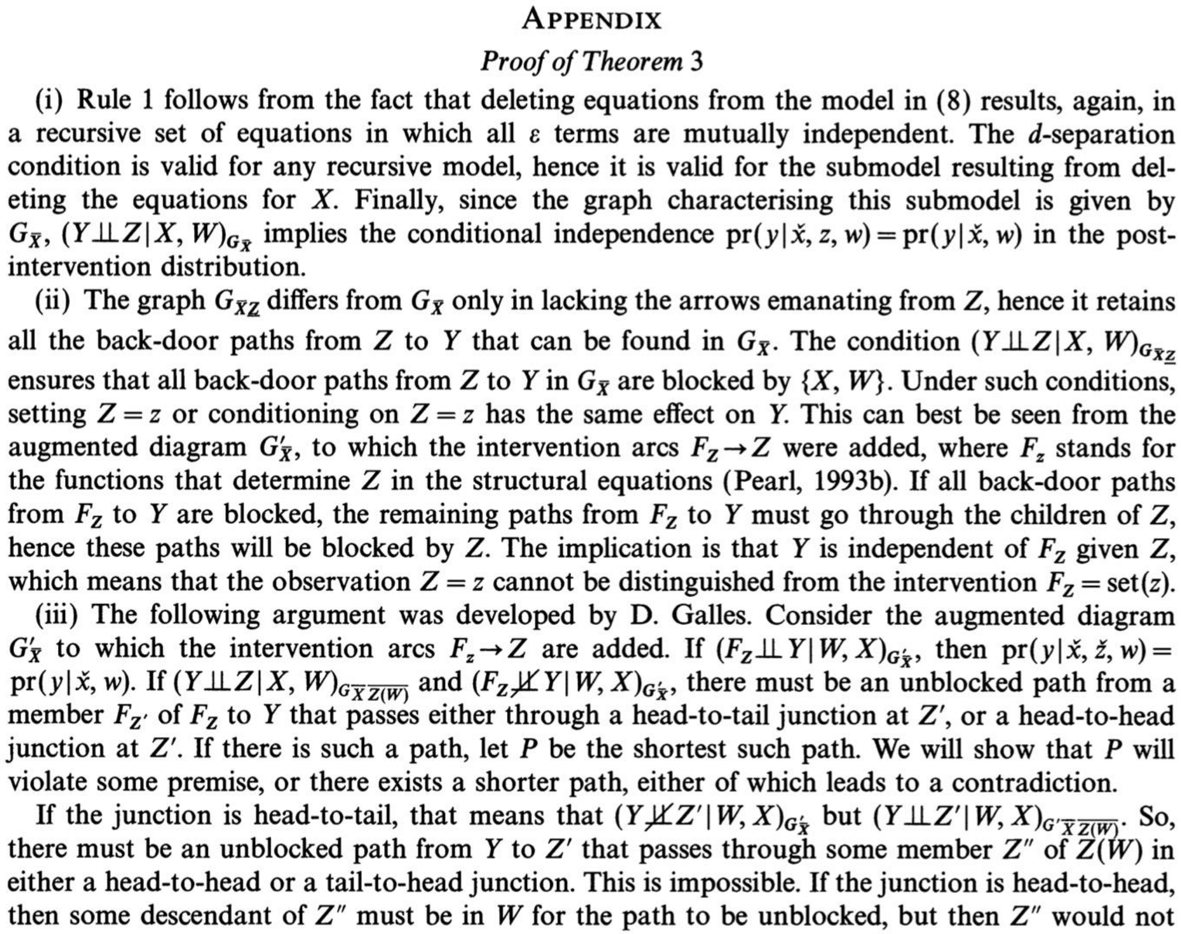
Rule 3 (*insertion/deletion of actions*):

$$\mathrm{pr}(y \mid \check{x}, \check{z}, w) = \mathrm{pr}(y \mid \check{x}, w) \quad \textit{if } (Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X},\, \overline{Z(W)}}}, \tag{12}$$

where Z(W) is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.

Pearl, Judea. "Causal diagrams for empirical research." Biometrika 82.4 (1995): 669-688

History of causal inference (Diagram)



Appendix

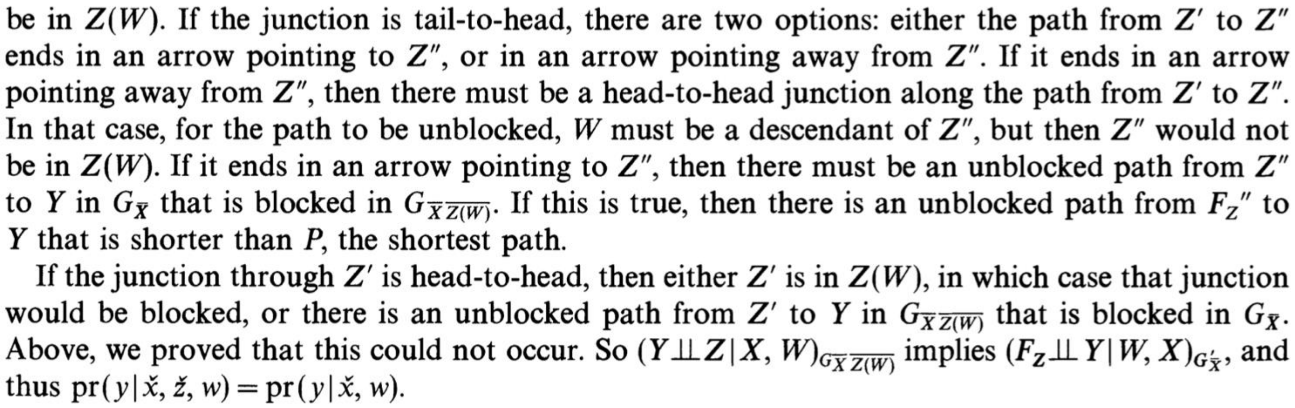
Proof of Theorem 3

(i) Rule 1 follows from the fact that deleting equations from the model in (8) results, again, in a recursive set of equations in which all ε terms are mutually independent. The d-separation condition is valid for any recursive model, hence it is valid for the submodel resulting from deleting the equations for X. Finally, since the graph characterising this submodel is given by $G_{\overline{X}}$, $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}}}$ implies the conditional independence $\mathrm{pr}(y \mid \check{x}, z, w) = \mathrm{pr}(y \mid \check{x}, w)$ in the post-intervention distribution.

(ii) The graph $G_{\overline{X}\underline{Z}}$ differs from $G_{\overline{X}}$ only in lacking the arrows emanating from Z, hence it retains all the back-door paths from Z to Y that can be found in $G_{\overline{X}}$. The condition $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\underline{Z}}}$ ensures that all back-door paths from Z to Y in $G_{\overline{X}}$ are blocked by $\{X, W\}$. Under such conditions, setting $Z = z$ or conditioning on $Z = z$ has the same effect on Y. This can best be seen from the augmented diagram $G'_{\overline{X}}$, to which the intervention arcs $F_Z \rightarrow Z$ were added, where F_z stands for the functions that determine Z in the structural equations (Pearl, 1993b). If all back-door paths from F_Z to Y are blocked, the remaining paths from F_Z to Y must go through the children of Z, hence these paths will be blocked by Z. The implication is that Y is independent of F_Z given Z, which means that the observation $Z = z$ cannot be distinguished from the intervention $F_Z = \mathrm{set}(z)$.

(iii) The following argument was developed by D. Galles. Consider the augmented diagram $G'_{\overline{X}}$ to which the intervention arcs $F_z \rightarrow Z$ are added. If $(F_Z \perp\!\!\!\perp Y \mid W, X)_{G'_{\overline{X}}}$, then $\mathrm{pr}(y \mid \check{x}, \check{z}, w) = \mathrm{pr}(y \mid \check{x}, w)$. If $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\,\overline{Z(W)}}}$ and $(F_Z \not\!\perp\!\!\!\perp Y \mid W, X)_{G'_{\overline{X}}}$, there must be an unblocked path from a member $F_{Z'}$ of F_Z to Y that passes either through a head-to-tail junction at Z', or a head-to-head junction at Z'. If there is such a path, let P be the shortest such path. We will show that P will violate some premise, or there exists a shorter path, either of which leads to a contradiction.

If the junction is head-to-tail, that means that $(Y \not\!\perp\!\!\!\perp Z' \mid W, X)_{G'_{\overline{X}}}$ but $(Y \perp\!\!\!\perp Z' \mid W, X)_{G'_{\overline{X}\,\overline{Z(W)}}}$. So, there must be an unblocked path from Y to Z' that passes through some member Z'' of $Z(W)$ in either a head-to-head or a tail-to-head junction. This is impossible. If the junction is head-to-head, then some descendant of Z'' must be in W for the path to be unblocked, but then Z'' would not



be in $Z(W)$. If the junction is tail-to-head, there are two options: either the path from Z' to Z'' ends in an arrow pointing to Z'', or in an arrow pointing away from Z''. If it ends in an arrow pointing away from Z'', then there must be a head-to-head junction along the path from Z' to Z''. In that case, for the path to be unblocked, W must be a descendant of Z'', but then Z'' would not be in $Z(W)$. If it ends in an arrow pointing to Z'', then there must be an unblocked path from Z'' to Y in $G_{\overline{X}}$ that is blocked in $G_{\overline{X}\,\overline{Z(W)}}$. If this is true, then there is an unblocked path from F_Z'' to Y that is shorter than P, the shortest path.

If the junction through Z' is head-to-head, then either Z' is in $Z(W)$, in which case that junction would be blocked, or there is an unblocked path from Z' to Y in $G_{\overline{X}\,\overline{Z(W)}}$ that is blocked in $G_{\overline{X}}$. Above, we proved that this could not occur. So $(Y \perp\!\!\!\perp Z \mid X, W)_{G_{\overline{X}\,\overline{Z(W)}}}$ implies $(F_Z \perp\!\!\!\perp Y \mid W, X)_{G'_{\overline{X}}}$, and thus $\mathrm{pr}(y \mid \check{x}, \check{z}, w) = \mathrm{pr}(y \mid \check{x}, w)$.

Pearl, Judea. "Causal diagrams for empirical research." Biometrika 82.4 (1995): 669-688