Lecture 6: Sequential Decisions

From One Decision to Many

DATA 8020: Advanced Causal Inference

Jin-Hong Du

Institute of Data Science, The University of Hong Kong

October 9, 2026

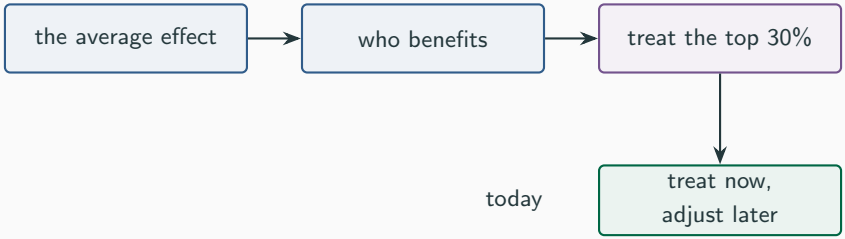
Today: what should we do, once and then over time?

By the end, you should be able to:

- evaluate and learn a treatment rule from logged data;
- explain why treatments over time need g-methods; and
- place a new method on the map of the field.

A rule can only be learned where past decisions explored.

Lecture 5 stopped one rung short of acting over time



Today the target is a **rule**, and then a rule that acts more than once.

The transplant registry used today is constructed; it contains no real patients.

One decision

A rule has a value: one number

$$V(\pi) = \mathbb{E}\big[\, Y\big(\pi(X)\big)\big], \quad \pi : X \mapsto \text{treat or not}$$

Regret: how far a rule falls below the best rule in its class.

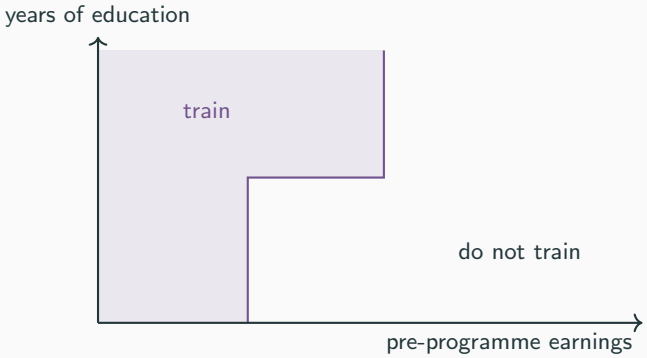
Evaluate a new rule with Lecture 4's score

$$\widehat{V}(\pi) = \tfrac{1}{n} \textstyle\sum_i \left[\pi(X_i)\,\Gamma_i(1) + \{1 - \pi(X_i)\}\,\Gamma_i(0)\right]$$

Same doubly robust score; machine learning calls this **off-policy evaluation**.

Dudík et al. (2014); Athey and Wager (2021)

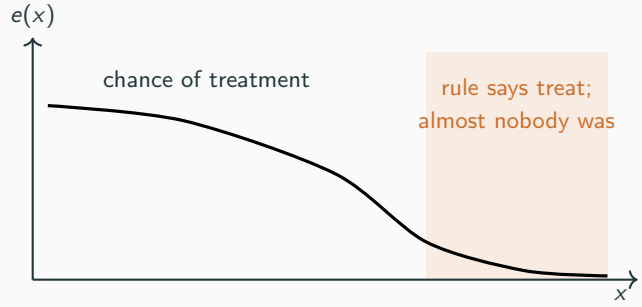
Choose the best rule within a simple class



Regret depends on the **class of rules**, not on how well you estimate $\tau(x)$.

JTPA job training, 9,223 applicants; region illustrative. Kitagawa and Tetenov (2018); Athey and Wager (2021)

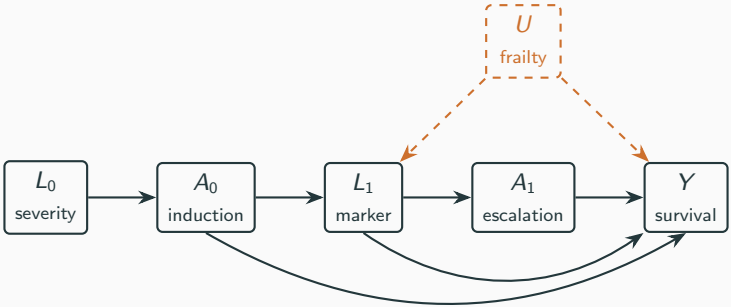
A rule nobody followed cannot be evaluated



Where $e(x) \approx 0$, the weight $1/e(x)$ explodes.

Over time

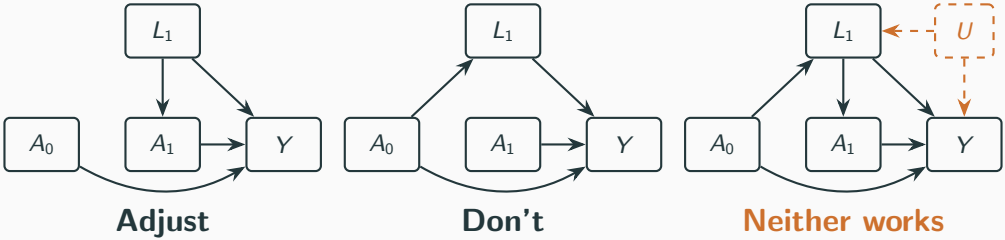
Two decisions, one timeline



Induction lowers the marker; a high marker prompts escalation.

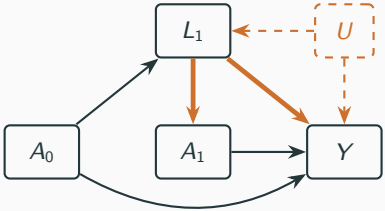
Constructed registry; true effect of always vs never treating: +2.0 years.

Three structures, three verdicts on adjusting for L_1

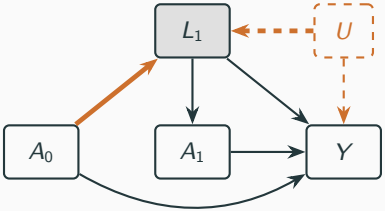


Left: L_1 only confounds A_1; middle: L_1 is only caused by A_0; right: both, plus hidden U.

Case 3: regression is wrong whichever way you choose



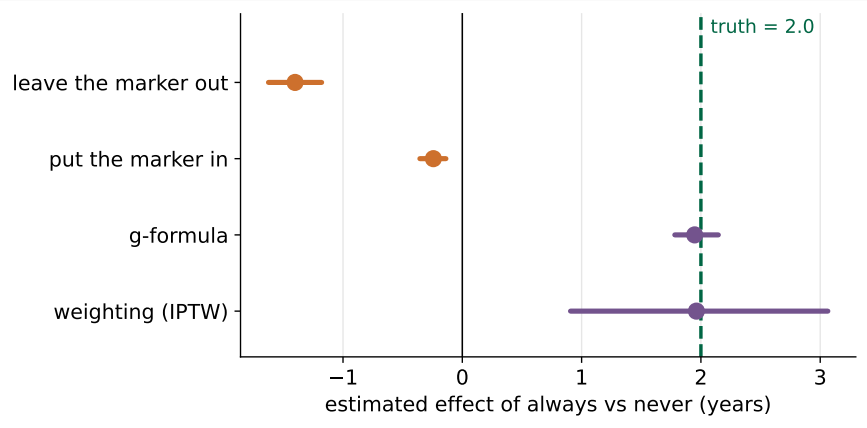
leave L_1 out: escalation confounded



put L_1 in: effect blocked, U let in

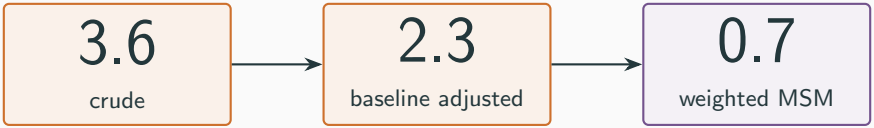
Adjusting for everything you measured is not safe.

In the registry, both regressions get the sign wrong



Constructed registry, $n = 4{,}000$; bars are 95% bootstrap intervals.

Real data: zidovudine and CD4 count in HIV



Mortality rate ratio for zidovudine: low CD4 prompts treatment, and treatment raises CD4.

Hernán et al. (2000); MSM interval 0.6–1.0.

Technical core 1: two assumptions give the g-formula

Sequential exchangeability: $Y(\bar{a}) \perp\!\!\!\perp A_t \mid$ past treatments and covariates

Sequential positivity: every treatment possible given the past

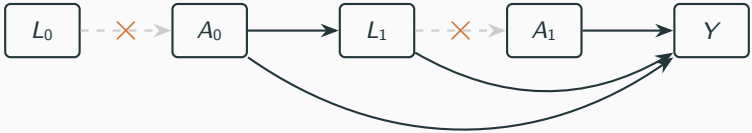
$$\mathbb{E}\{Y(a_0, a_1)\} = \sum_{l_0, l_1} \mathbb{E}(Y \mid l_0, a_0, l_1, a_1)\; p(l_1 \mid l_0, a_0)\; p(l_0)$$

Average L_1 over what a_0 **would produce**, not what was observed.

Robins (1986); derivation in the notes.

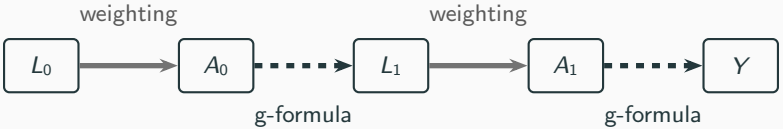
Technical core 2: or reweight, and the arrow into treatment disappears

$$W = \prod_t \frac{1}{\mathbb{P}(A_t \mid \text{past})}$$



Marginal structural models: Robins et al. (2000).

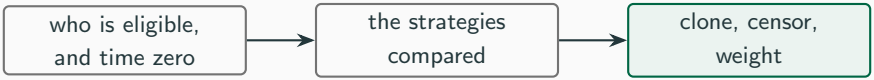
Which arrows does each method model?



Model both sets of arrows ⇒ doubly robust.

Bang and Robins (2005); g-estimation: Robins (1994); textbook: Hernán and Robins (2020, Part III).

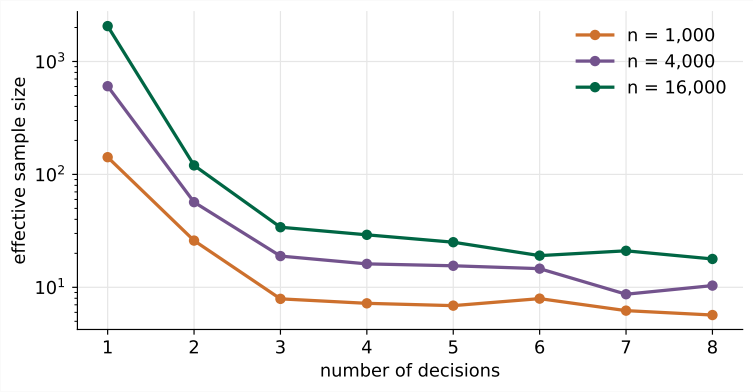
Before analysing, write down the trial you wish you had



Target-trial emulation makes the g-methods' assumptions visible.

Hernán and Robins (2016)

Overlap is spent at every decision

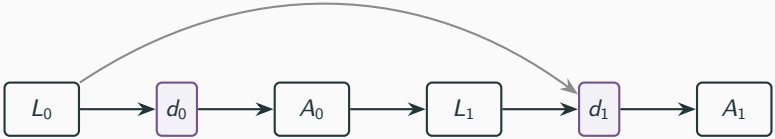


After three decisions, 16× the patients buys only about 4× the effective sample.

Constructed registry extended to K visits; true propensities; weights for "always treat".

Dynamic regimes

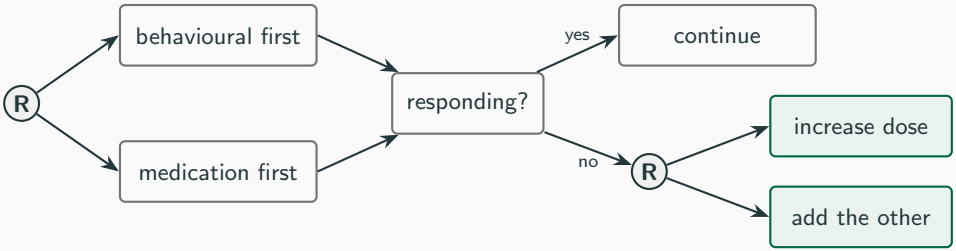
A regime is a rule that reads the history



"Escalate if the marker stays high" is a rule, not a fixed plan.

Murphy (2003); review: Chakraborty and Murphy (2014).

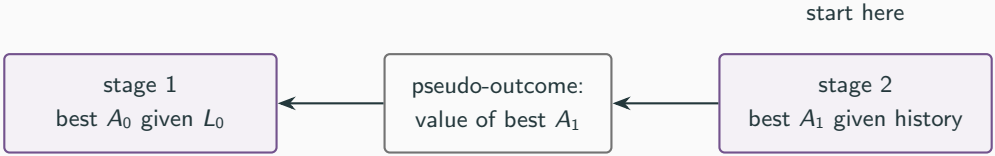
The experiment for a regime randomizes twice



A SMART: 146 children with ADHD, randomized twice.

Pelham et al. (2016); design: Murphy (2005).

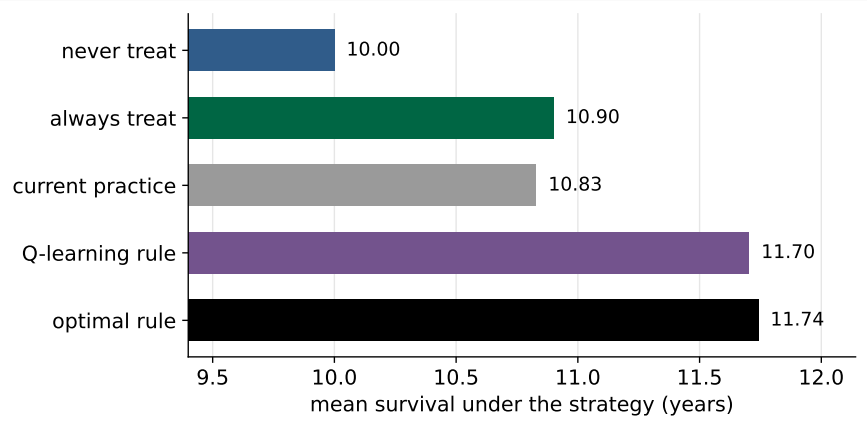
Learn the regime backwards: Q-learning



Each stage is a regression; the last decision is solved first.

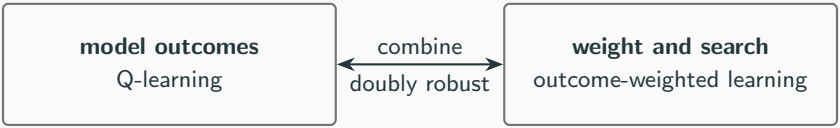
Murphy (2003); same recursion as Watkins and Dayan (1992).

A rule that reads the patient beats every fixed plan



Constructed registry with tailored effects; values simulated from the true mechanism.

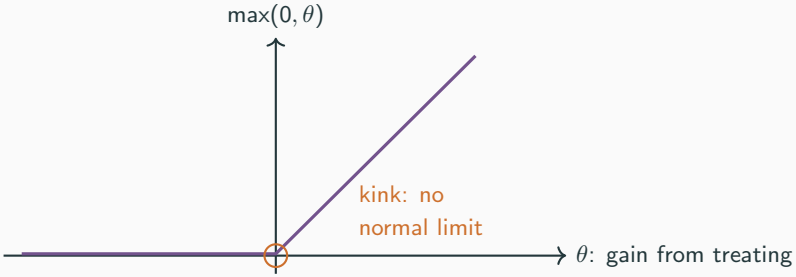
Model the outcome, or search the rules directly



The same trade-off as for one decision.

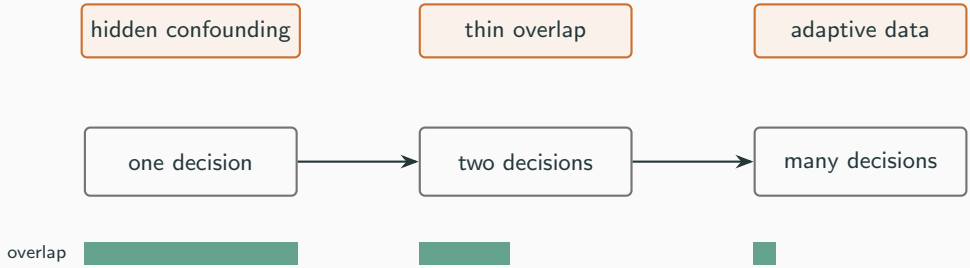
Zhao et al. (2012); Zhang et al. (2012)

Inference breaks where treatment makes no difference



Laber et al. (2014); Luedtke and van der Laan (2016)

Half I in one picture



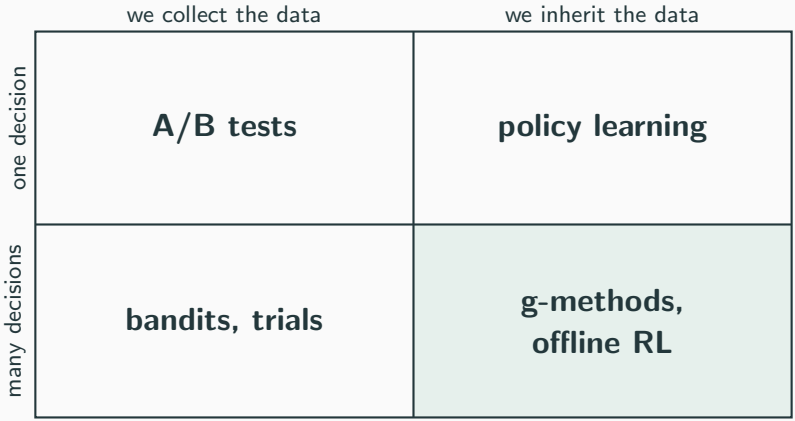
After the break: what the field is doing about each orange box.

Break

10 minutes

Field map

Four corners of one field



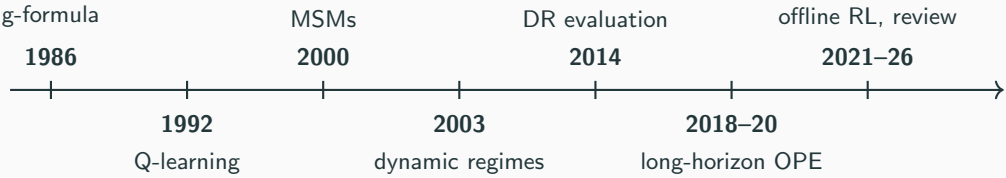
Shaded: where Half I lived.

Same problem, two vocabularies

causal inference		reinforcement learning
regime	⟷	policy
g-formula	⟷	model-based evaluation
weighting	⟷	importance sampling
positivity	⟷	coverage

RL usually adds one assumption: the current state summarizes the past.

Forty years in seven ticks

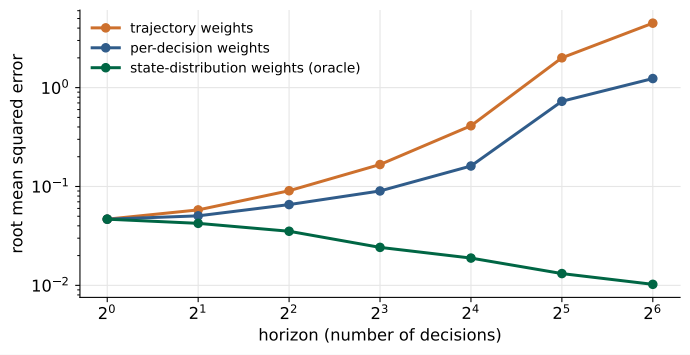


Full dated list with references in the notes.

Frontier

Long horizons: can the weights stop compounding?

relaxes: overlap spent at every decision

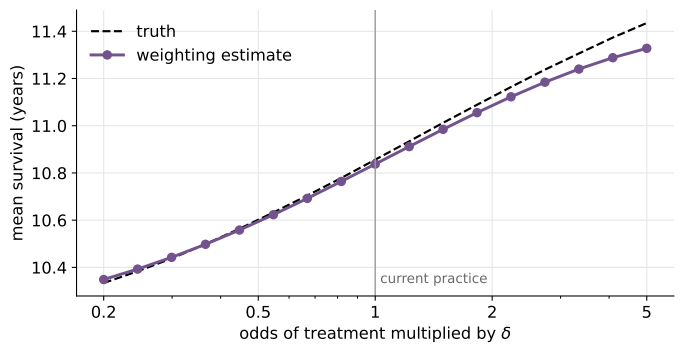


Under Markov structure, reweight states, not whole trajectories.

Liu et al. (2018); Kallus and Uehara (2020); constructed process; green curve uses the true ratio.

Thin overlap: can we ask a smaller question?

relaxes: positivity

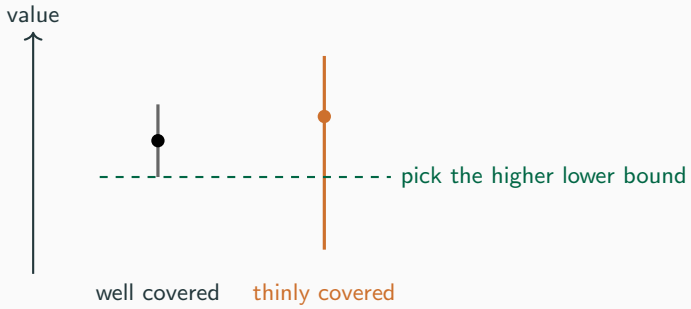


Multiply everyone's odds of treatment by δ, instead of forcing "always".

Kennedy (2019); Díaz et al. (2023); constructed registry.

Thin overlap: can we refuse to trust thin data?

relaxes: coverage of the logged data

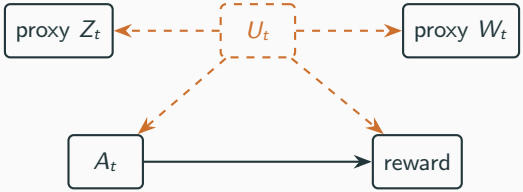


Pessimism: the higher point estimate loses.

Jin et al. (2021); tutorial: Levine et al. (2020).

Hidden confounding: can proxies stand in for the hidden state?

relaxes: sequential exchangeability

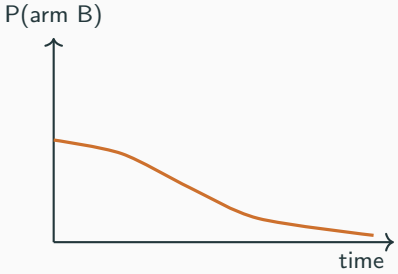


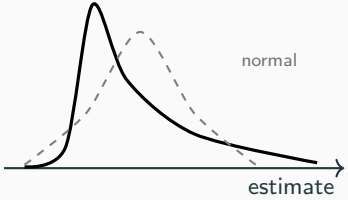
Proximal RL, or bounds under a sensitivity model; Week 9 has the static version.

Tennenholtz et al. (2020); Bennett and Kallus (2024); Kallus and Zhou (2020).

Adaptive data: is the usual interval still valid?

relaxes: independent sampling



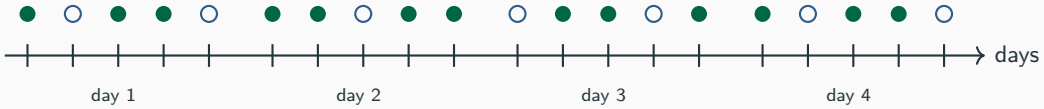


Not always: reweight adaptively, or use intervals valid at any stopping time.

Zhang et al. (2020); Hadad et al. (2021); Waudby-Smith et al. (2024); schematic.

Designed data: randomize many times a day

relaxes: nothing; design the data instead

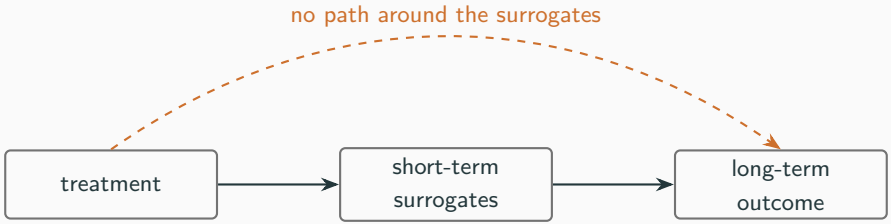


A micro-randomized trial: HeartSteps, 5 decision points a day, probability 0.6, 42 days.

Klasnja et al. (2015); Klasnja et al. (2019); filled: suggestion sent; pattern illustrative.

Late outcomes: can short-term proxies speak for them?

relaxes: observing the outcome of the decision



A surrogate index predicts the long-term outcome from several short ones.

Athey et al. (2026)

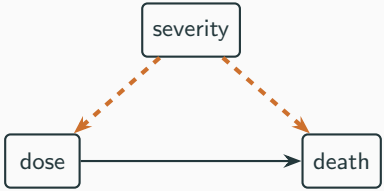
Audit: a reinforcement-learning policy for sepsis

A policy learned from ICU records recommends fluids and vasopressors;
mortality was lowest when clinicians' doses matched it.

1. Which arrow could fool this analysis?
2. Where did clinicians rarely do what the policy recommends?

Pairs, 7 minutes. Komorowski et al. (2018)

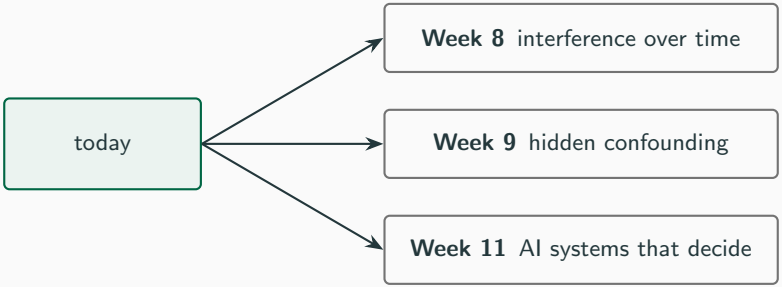
Audit: what to check



1. Sicker patients get more treatment and die more: confounding by indication.
2. Where the policy departs from practice, few patients carry the estimate.

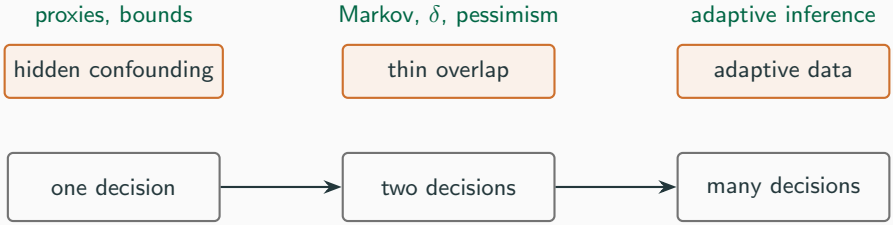
Gottesman et al. (2019), written with members of the original team.

Where today returns



Take stock

Overlap is the currency



Learn a rule only where past decisions explored.

Next

- Lecture 7, October 23: interference and experiment design.
- Reading week: Chakraborty and Murphy (2014); Uehara et al. (2026).
- Notebook: the sign flip, overlap decay, Q-learning.

References i

Athey, S., Chetty, R., Imbens, G. W., and Kang, H. (2026). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. *The Review of Economic Studies*, 93(4):2284–2312.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

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

Bennett, A. and Kallus, N. (2024). Proximal reinforcement learning: Efficient off-policy evaluation in partially observed Markov decision processes. *Operations Research*, 72(3):1071–1086.

Chakraborty, B. and Murphy, S. A. (2014). Dynamic treatment regimes. *Annual Review of Statistics and Its Application*, 1:447–464.

Díaz, I., Williams, N., Hoffman, K. L., and Schenck, E. J. (2023). Nonparametric causal effects based on longitudinal modified treatment policies. *Journal of the American Statistical Association*, 118(542):846–857.

Dudík, M., Erhan, D., Langford, J., and Li, L. (2014). Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and Celi, L. A. (2019). Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–18.

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15):e2014602118.

Hernán, M. Á., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5):561–570.

Hernán, M. A. and Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758–764.

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

Jin, Y., Yang, Z., and Wang, Z. (2021). Is pessimism provably efficient for offline RL? In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5084–5096.

Kallus, N. and Uehara, M. (2020). Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63.

References ii

Kallus, N. and Zhou, A. (2020). Confounding-robust policy evaluation in infinite-horizon reinforcement learning. In *Advances in Neural Information Processing Systems 33*.

Kennedy, E. H. (2019). Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656.

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616.

Klasnja, P., Hekler, E. B., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., and Murphy, S. A. (2015). Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(Suppl):1220–1228.

Klasnja, P., Smith, S., Seewald, N. J., Lee, A., Hall, K., Luers, B., Hekler, E. B., and Murphy, S. A. (2019). Efficacy of contextually tailored suggestions for physical activity: A micro-randomized optimization trial of HeartSteps. *Annals of Behavioral Medicine*, 53(6):573–582.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720.

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., and Murphy, S. A. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electronic Journal of Statistics*, 8(1):1225–1272.

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems.

Liu, Q., Li, L., Tang, Z., and Zhou, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems 31*.

Luedtke, A. R. and van der Laan, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics*, 44(2):713–742.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24(10):1455–1481.

Pelham, Jr., W. E., Fabiano, G. A., Waxmonsky, J. G., Greiner, A. R., Gnagy, E. M., Pelham, III, W. E., Coxe, S., Verley, J., Bhatia, I., Hart, K., Karch, K., Konijnendijk, E., Tresco, K., Nahum-Shani, I., and Murphy, S. A. (2016). Treatment sequencing for childhood ADHD: A multiple-randomization study of adaptive medication and behavioral interventions. *Journal of Clinical Child & Adolescent Psychology*, 45(4):396–415.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9–12):1393–1512.

References iii

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics—Theory and Methods*, 23(8):2379–2412.

Robins, J. M., Hernán, M. Á., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Tennenholtz, G., Shalit, U., and Mannor, S. (2020). Off-policy evaluation in partially observable environments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(06):10276–10283.

Uehara, M., Shi, C., and Kallus, N. (2026). A review of off-policy evaluation in reinforcement learning. *Statistical Science*, 41(3):561–581.

Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3–4):279–292.

Waudby-Smith, I., Wu, L., Ramdas, A., Karampatziakis, N., and Mineiro, P. (2024). Anytime-valid off-policy inference for contextual bandits. *ACM / IMS Journal of Data Science*, 1(3):1–42.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018.

Zhang, K. W., Janson, L., and Murphy, S. A. (2020). Inference for batched bandits. In *Advances in Neural Information Processing Systems 33*.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118.

Questions?

Notebook: the sign flip, overlap decay, and a learned regime