

Machine Learning for Healthcare

6.871, HST.956

Lecture 22: Dynamic treatment regimes & off-policy reinforcement learning

David Sontag



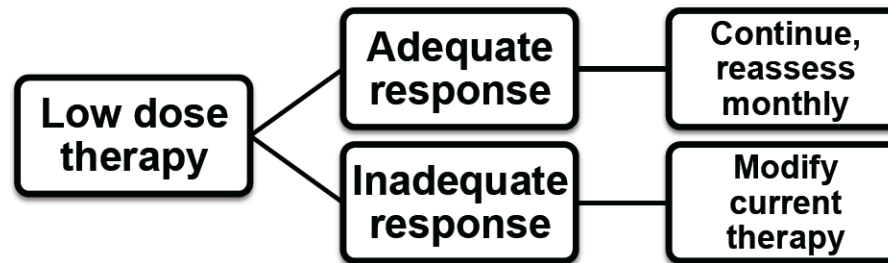
A path to personalized medicine

- Clinical practice: Clinicians make (a series of) treatment decision(s) over the course of a patient's disease or disorder
 - Key decision points in the disease process
 - Could be a fixed schedule, a milestone in the disease process, or an event necessitating a decision
 - Several treatment options at each decision point
- Thus: treatment in practice involves **sequential decision-making** based on accruing information

Dynamic treatment regime

- Sequential decision rules, each corresponding to a key decision point
- Each rule tells us treatment to be given from among the available options based on the accrued information on the patient to that point
- Taken together, the rules define an algorithm for making treatment decisions
- *Dynamic* because the treatment action can vary depending on the accrued information



Example: ADHD therapy



- Decision 1: Low-dose therapy – 2 options: medication or behavior modification
- Subsequent monthly decisions:
 - Responders: Continue initial therapy
 - Non-responders – 2 options: add the other therapy or increase dose of current therapy
- Objective: maximize *end-of-school-year performance*

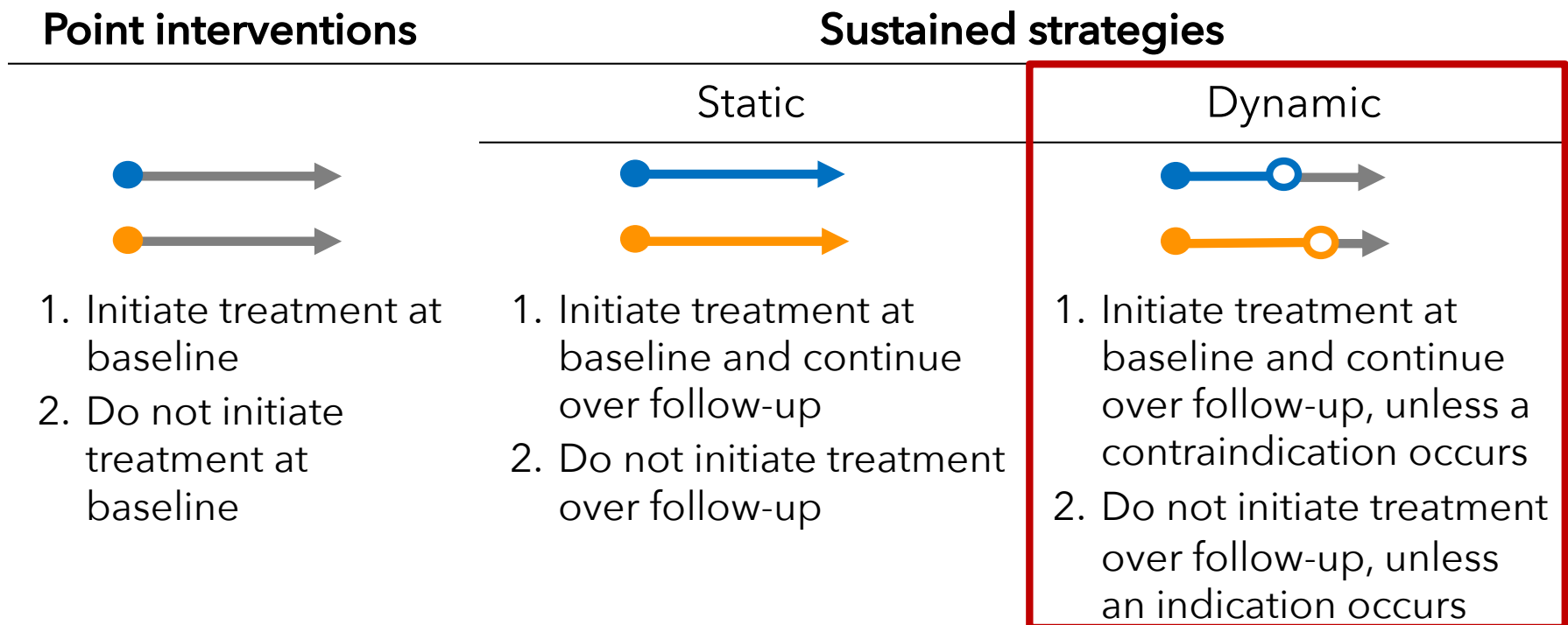
(Material from Marie Davidian, *An Introduction to Dynamic Treatment Regimes*; example from Susan Murphy)

Example: Physical activity for men with prostate cancer

- **Treatment regimes:** Initiate 1 of 6 physical activity strategies at baseline and continue it over follow-up until development of a condition limiting physical activity
- **(Vigorous activity)** Regime 1: 1.25 hrs/wk; Regime 2: 2.5 hrs/wk; Regime 3: 3.75 hrs/wk 
- **(Moderate activity)** Regime 4: 2.5 hrs/wk; Regime 5: 5.0 hrs/wk; Regime 6: 7.5 hrs/wk 
- Outcome: all-cause mortality within 10 years of diagnosis

Example: Physical activity for men with prostate cancer

- This is a dynamic treatment strategy because of the decision when to stop



(Dickerman et al., Guideline-Based Physical Activity and Survival Among US Men With Nonmetastatic Prostate Cancer. Am J Epidemiol. 2019)

Example: First-line treatment for multiple myeloma

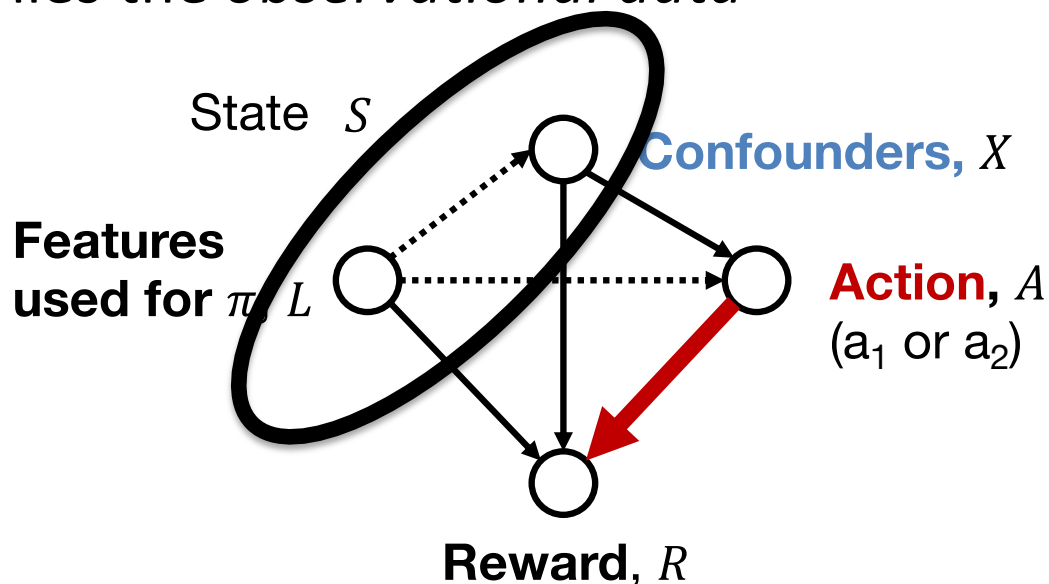
- Decision 1: Induction chemotherapy (options C_1, C_2)
- Decision 2:
 - Maintenance treatment for patients who *respond* (options M_1, M_2)
 - Start a different cancer treatment for those who don't respond (options S_1, S_2)
- Objective: maximize *survival time*
- Example rules for decision 1:
 - C_1 : If “age < 65 and in excellent physical health”, give bortezomib, lenalidomide, dexamethasone chemotherapy followed by autologous stem cell transplant. Otherwise, treat with daratumumab, bortezomib, melphalan, & prednisone.
 - C_2 : treat everyone with daratumumab, bortezomib, melphalan, & prednisone

Example: First-line treatment for multiple myeloma

- Which is the best treatment regime (policy)?
- Evaluate each of the following 8 dynamic regimes:
 1. Give C_1 followed by (M_1 if response, S_1 if no response)
 2. Give C_1 followed by (M_1 if response, S_2 if no response)
 3. Give C_1 followed by (M_2 if response, S_1 if no response)
 4. Give C_1 followed by (M_2 if response, S_2 if no response)
 5. Give C_2 followed by (M_1 if response, S_1 if no response)
 6. Give C_2 followed by (M_1 if response, S_2 if no response)
 7. Give C_2 followed by (M_2 if response, S_1 if no response)
 8. Give C_2 followed by (M_2 if response, S_2 if no response)
- Goal: evaluate the average *outcome* if all patients in the population were to follow each regime

Warm up: policies for point interventions (also, static policies)

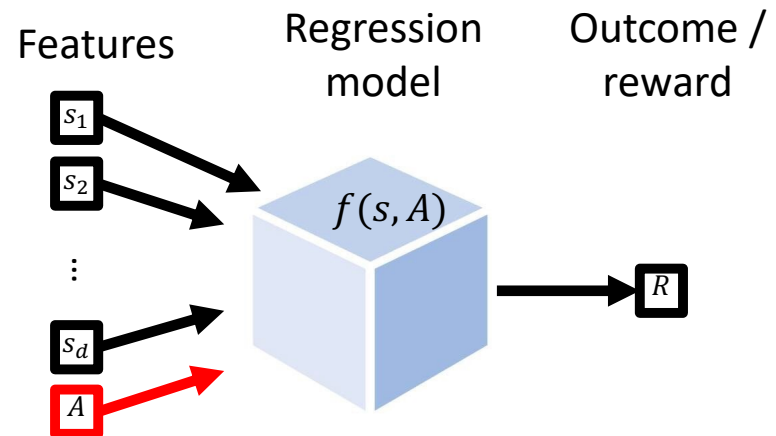
- Suppose someone gave us a policy $\pi(l)$ that outputs a_1 vs a_2
- How do we evaluate it?
- In Lecture 12, we gave two approaches, one based on potential outcomes and the other based on propensity scores
- In both cases, we have to first consider the causal graph that underlies the *observational data*



Switched notation to what's more typically used in RL
action A : Treatment T
reward R : Outcome Y

Evaluating policies using covariate adjustment (from lecture 12)

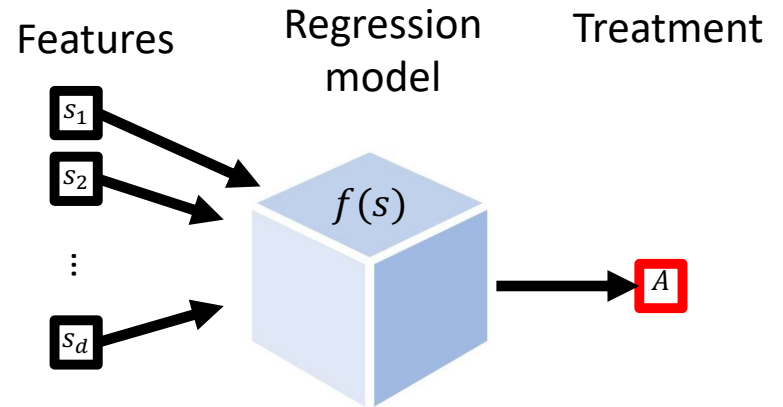
- First, use machine learning to obtain a model that can predict potential outcomes (we need ignorability, overlap)
- Then, use this model to estimate average reward of actions this policy would take:



$$\hat{Q}(\pi) = \frac{1}{n} \sum_{i=1}^n f(\underbrace{l_i, x_i}_{s_i}, \pi(l_i))$$

Evaluating policies using inverse propensity scores (from lecture 12)

- First, use machine learning to obtain $\hat{p}(A|s) = f(s)$, estimated propensity scores



- Then, use this model to reweight the observed rewards, accounting for dataset shift from observational policy to policy we wish to evaluate:

$$\hat{Q}^{IPW}(\pi) = \frac{1}{n} \sum_{i=1}^n \frac{1[a_i = \pi(l_i)]}{\hat{p}(a_i | s_i)} R_i$$

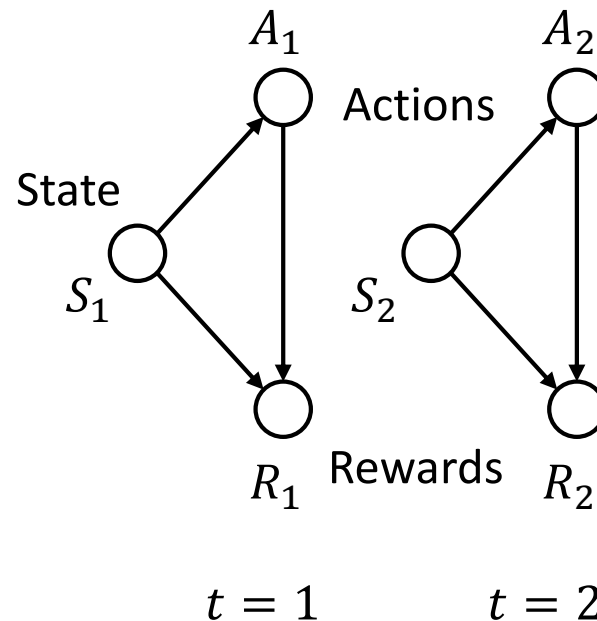
Causal graph for dynamic treatment regimes

- Consider the *true* causal graph that generated the sequential treatment decisions observed in the data
- Best case scenario: independent decisions!

Causal graph for ~~dynamic~~ treatment regimes

- Consider the *true* causal graph that generated the sequential treatment decisions observed in the data
- Best case scenario: independent decisions!

Very important:
 S_t includes *both* L_t
 (variables used for
 π_t) and X_t
 (confounders of
 clinician treatment
 decision A_t and
 current reward R_t)



... **Ignorability**
 $R_t(a) \perp\!\!\!\perp A_t \mid S_t$

At each time step, we get completely fresh information that impacts next treatment decision

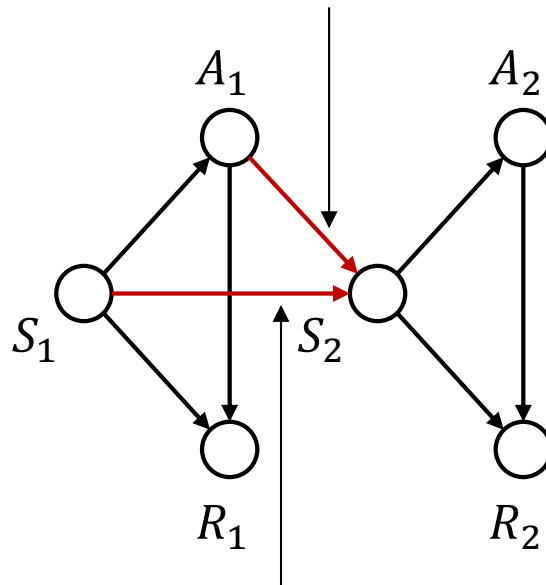
Causal graph for dynamic treatment regimes

- Consider the *true* causal graph that generated the sequential treatment decisions observed in the data

Very important:

To maintain ignorability, S_t should include *both* L_t (variables used for π_t) and X_t (confounders of clinician treatment decision A_t and current and future rewards R_t, R_{t+1}, \dots)

Anna's health status depends on how we treated her



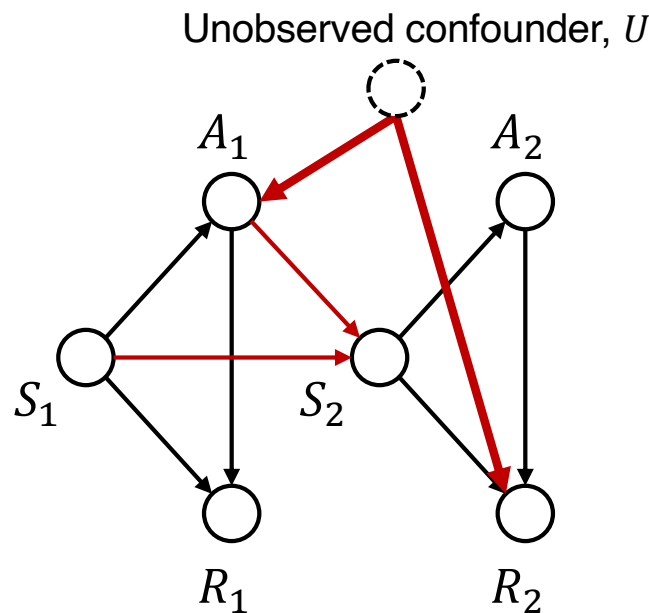
It is likely that if Anna is diabetic, she will remain so

Causal graph for dynamic treatment regimes

- Consider the *true* causal graph that generated the sequential treatment decisions observed in the data

Very important:

To maintain ignorability, S_t should include *both* L_t (variables used for π_t) and X_t (confounders of clinician treatment decision A_t and current and future rewards R_t, R_{t+1}, \dots)

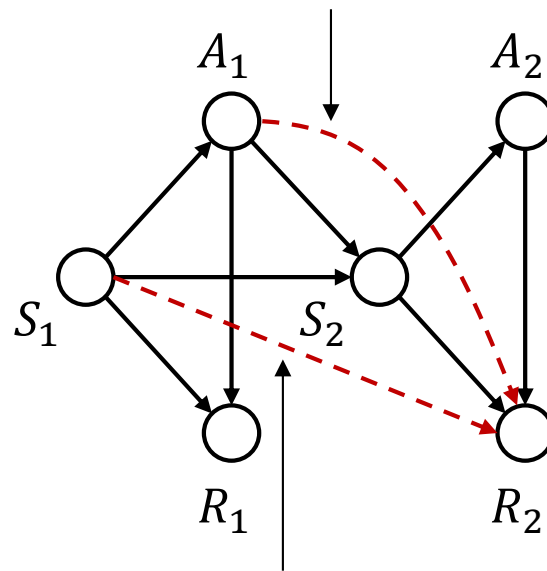


Ignorability violated

Causal graph for dynamic treatment regimes

- Consider the *true* causal graph that generated the sequential treatment decisions observed in the data

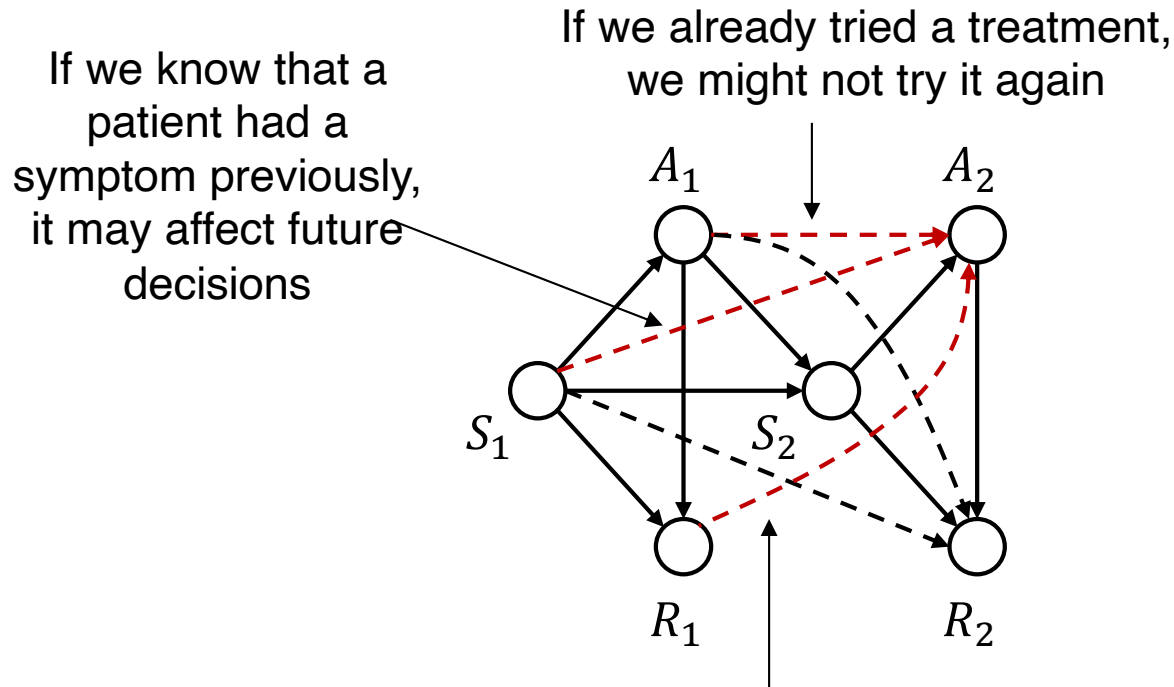
The outcome at a later time point may depend on earlier choices



The outcome at a later time may depend on an earlier state

Causal graph for dynamic treatment regimes

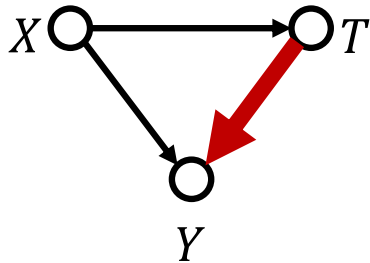
- Consider the *true* causal graph that generated the sequential treatment decisions observed in the data



TL;DR:
Think about possible short- and long-term confounders and include them in S

If the last treatment was unsuccessful, it may change our next choice

Assumptions for evaluation of dynamic treatment regimes



Single-step case

Strong ignorability:

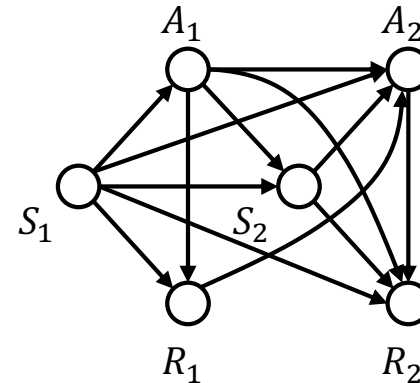
$$Y(0), Y(1) \perp\!\!\!\perp T \mid X$$

“No *hidden* confounders”

Overlap:

$$\forall x, t: p(T = t \mid X = x) > 0$$

“All actions possible”



Sequential case

Sequential randomization:

$$\forall t' \geq t: R_{t'} \perp\!\!\!\perp A_t \mid \bar{S}_t, \bar{A}_{t-1}$$



“Reward indep. of policy given history”

Positivity:

$$\forall a, t: p(A_t = a \mid \bar{S}_t, \bar{A}_{t-1}) > 0$$

“All actions possible at all times”

Physical activity for men with prostate cancer

- **Treatment regimes:** Initiate 1 of 6 physical activity strategies at baseline and continue it over follow-up until development of a condition limiting physical activity  vs. 
- **Outcome:** all-cause mortality within 10 years of diagnosis



What data do we need to collect?

Web Table 2. Covariates Used to Model 10-Year Risk of All-Cause Mortality Among Men With Nonmetastatic Prostate Cancer, Health Professionals Follow-Up Study.

A. Time-fixed covariates	Functional form as predictor	Variable name	Categories
Baseline (assessed in first post-diagnostic questionnaire)			
Age	4 categories	baseage_1	<65 years
		baseage_2	65-69.9 years
		baseage_3	70-74.9 years
		baseage_4	≥75 years
Clinical stage at diagnosis	2 categories	stage_1	T1
		stage_2	T2, T3, T4, N1/M0
Prostate-specific antigen level at diagnosis	2 categories	psa_1	<4 ng/mL
		psa_2	≥4 ng/mL
Gleason grade at diagnosis	3 categories	gleason_1	<7
		gleason_2	7
		gleason_3	>7
Primary treatment	3 categories	treat_1	Radical prostatectomy
		treat_2	Radiation
		treat_3	Hormones, watchful waiting, other
Parental history of myocardial infarction before age 60	Indicator	fhxmi	Yes/No

(Dickerman et al., Guideline-Based Physical Activity and Survival Among US Men With Nonmetastatic Prostate Cancer. Am J Epidemiol. 2019)

Physical activity for men with prostate cancer



- **Treatment regimes:** Initiate 1 of 6 physical activity strategies at baseline and continue it over follow-up until development of a condition limiting physical activity  vs. 
- **Outcome:** all-cause mortality within 10 years of diagnosis

What data do we need to collect?

Pre-baseline (assessed in first pre-diagnostic questionnaire)			
BMI	4 categories	bmi_pre_1	<18.5 kg/m ²
		bmi_pre_2	18.5-24.9 kg/m ²
		bmi_pre_3	25.0-29.9 kg/m ²
		bmi_pre_4	≥30 kg/m ²
Vigorous physical activity	4 categories	vigact_pre_1	<1.25 hour/week
		vigact_pre_2	1.25-2.49 hours/week
		vigact_pre_3	2.50-3.74 hours/week
		vigact_pre_4	≥3.75 hours/week
Moderate physical activity	4 categories	modact_pre_1	<2.5 hours/week
		modact_pre_2	2.5-4.9 hours/week
		modact_pre_3	5-7.4 hours/week
		modact_pre_4	≥7.5 hours/week
Smoking history	Indicator	smkhx	Yes/No

(Dickerman et al., Guideline-Based Physical Activity and Survival Among US Men With Nonmetastatic Prostate Cancer. Am J Epidemiol. 2019)

Physical activity for men with prostate cancer

- **Treatment regimes:** Initiate 1 of 6 physical activity strategies at baseline and continue it over follow-up until development of a condition limiting physical activity  vs. 
- **Outcome:** all-cause mortality within 10 years of diagnosis

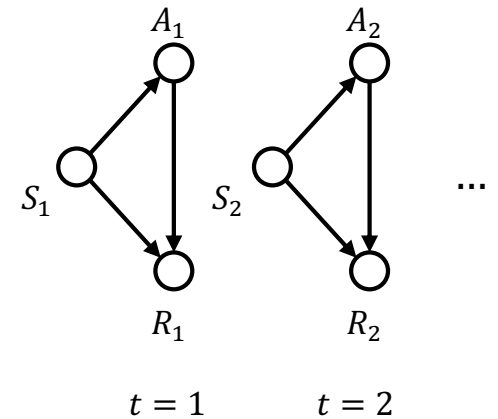
What data do we need to collect?

B. Time-varying covariates ^a	Modeling as dependent	Variable name	Functional form as predictor	Category or knot locations
Period of follow-up	Not predicted	period	5 period indicators	N/A
BMI	Linear (on log scale) ^b	bmi	4 categories	18.5, 25, 30 kg/m ²
Vigorous physical activity	Logistic, then log-linear ^c	vigact	Restricted cubic splines, 3 knots	1.25, 2.5, 3.75 hours/week
Moderate physical activity	Linear ^b	modcat	Restricted cubic splines, 3 knots	2.5, 5, 7.5 hours/week
Development of functional impairment, metastasis, myocardial infarction, stroke, congestive heart failure, or amyotrophic lateral sclerosis	Logistic to failure ^d	xcond	Indicator and time since switch	N/A

(Dickerman et al., Guideline-Based Physical Activity and Survival Among US Men With Nonmetastatic Prostate Cancer. Am J Epidemiol. 2019)

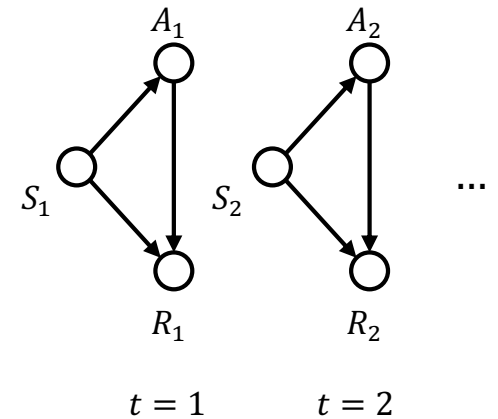
Warm up: Evaluating dynamic treatment regimes

- As a warmup, consider the simplified causal model shown on the right
- Assume that the policy we are evaluating, π , is given by a different rule π_t for each time step t



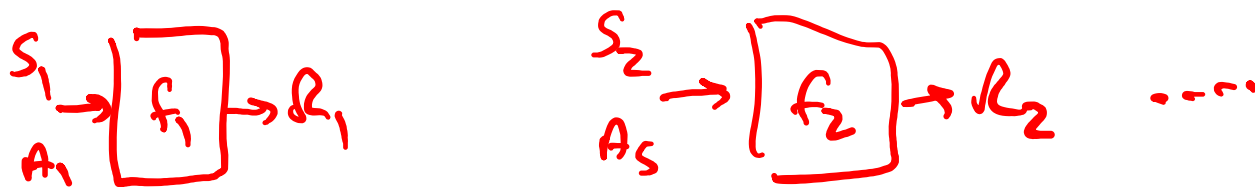
Warm up: Evaluating dynamic treatment regimes

- As a warmup, consider the simplified causal model shown on the right
- Assume that the policy we are evaluating, π , is given by a different rule π_t for each time step t



$$Q(\pi) = \mathbb{E}_{S_1, R_1(A_1), S_2, R_2(A_2), \dots} \left[\sum_t R_t \right] \left. \vphantom{\mathbb{E}} \right\} \text{adjustment formula}$$

$$\Rightarrow \sum_t \mathbb{E}_{S_t} [R_t | S_t, \pi_t(S_t)] \leftarrow$$

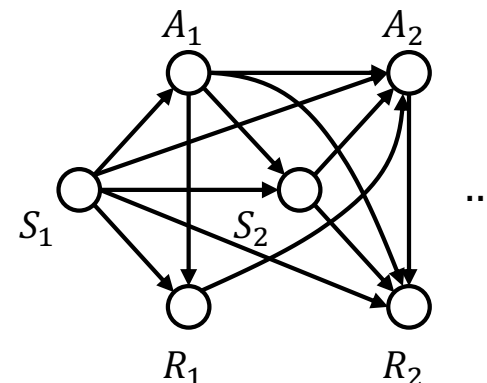


$$\hat{Q}(\pi) = \frac{1}{n} \sum_i \sum_t f(S_t, \pi_t(S_t))$$

T regressions

Evaluating dynamic treatment regimes

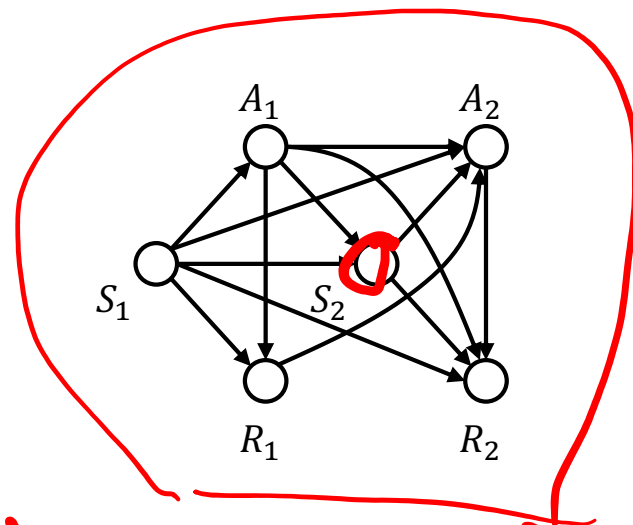
- Notice that the same estimator *does not* make sense when, e.g., S_2 depends on A_1
- The distribution of states S_2 will be affected by the policy's choice of actions A_1
 - Cannot use the observational distribution



Evaluating dynamic treatment regimes

$$Q(\pi) = \mathbb{E}_P \left[\sum_{t=1}^T R_t \right]$$

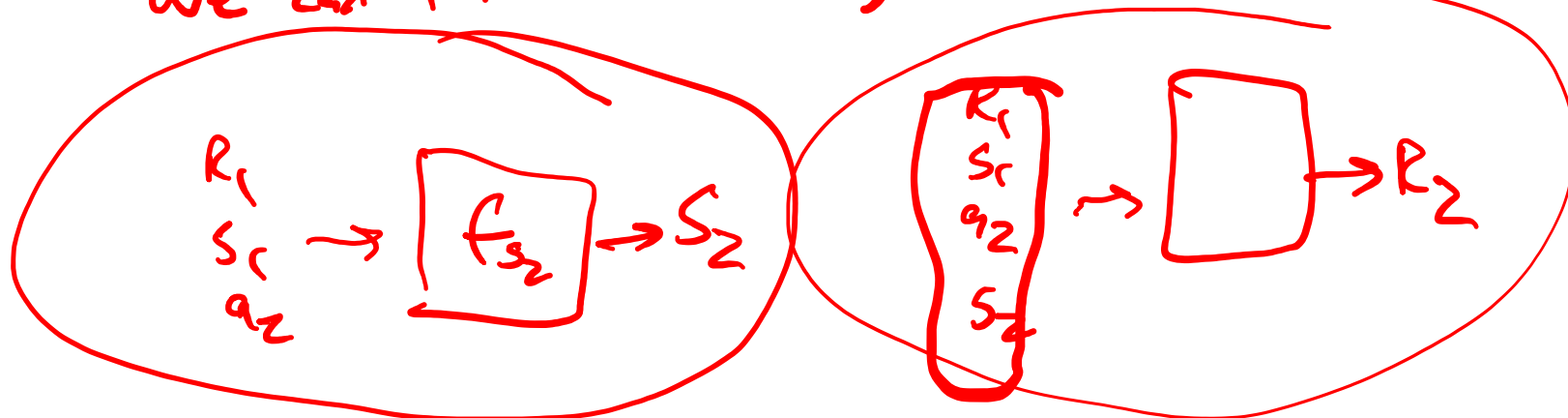
$$P(s_1, R_1(\pi(s_1)), s_2, R_2(\pi(s_1), \pi(s_2)), \dots)$$



$$= P(s_1) P(R_1 | s_1, a_1 = \pi_1(s_1)) \cdot$$

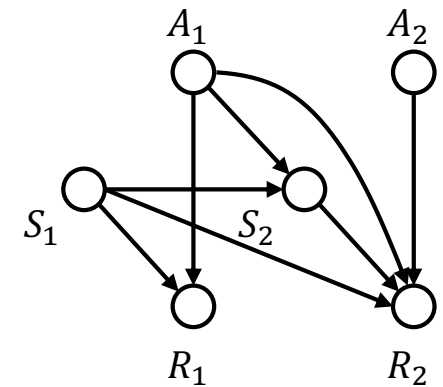
$$P(s_2 | R_1, s_1, a_2 = \pi_2(s_2)) P(R_2 | R_1, s_1, a_2 = \pi_2(s_2), s_2)$$

We can fit this using our observational data



Evaluating dynamic treatment regimes: parametric G-formula

- ① **Fit parametric regression models** for confounders and death at each follow-up time t as a function of treatment and covariate history among those under follow-up at time t
- ② **Monte Carlo simulation** to generate a 10,000-person population under each strategy by sampling with replacement from the original study population (to estimate the standardized cumulative risk under a given strategy)
- ③ **Repeat in 500 bootstrap samples** to obtain 95% confidence intervals (CIs)





Concern: Errors may compound; also, may be insufficient data for any one time step.

[James Robins. 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*, 1986.]

For recent work, see: Rui Li et al., G-Net: a Recurrent Network Approach to G-Computation for Counterfactual Prediction Under a Dynamic Treatment Regime. *Proceedings of Machine Learning Research* 158:282–297, 2021.]

Physical activity for men with prostate cancer

- **Treatment regimes:** Initiate 1 of 6 physical activity strategies at baseline and continue it over follow-up until development of a condition limiting physical activity  vs. 
- **Outcome:** all-cause mortality within 10 years of diagnosis

What data do we need to collect?

B. Time-varying covariates ^a	Modeling as dependent	Variable name	Functional form as predictor	Category or knot locations
Period of follow-up	Not predicted	period	5 period indicators	N/A
BMI	Linear (on log scale) ^b	bmi	4 categories	18.5, 25, 30 kg/m ²
Vigorous physical activity	Logistic, then log-linear ^c	vigact	Restricted cubic splines, 3 knots	1.25, 2.5, 3.75 hours/week
Moderate physical activity	Linear ^b	modcat	Restricted cubic splines, 3 knots	2.5, 5, 7.5 hours/week
Development of functional impairment, metastasis, myocardial infarction, stroke, congestive heart failure, or amyotrophic lateral sclerosis	Logistic to failure ^d	xcond	Indicator and time since switch	N/A

(Dickerman et al., Guideline-Based Physical Activity and Survival Among US Men With Nonmetastatic Prostate Cancer. Am J Epidemiol. 2019)

Physical activity for men with prostate cancer

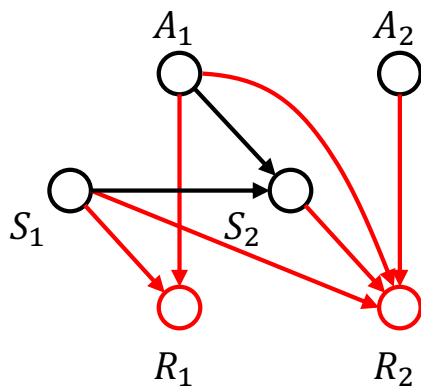
Model 1 Outcome model

The LOGISTIC Procedure

Model Information

Data Set	WORK.PARAM
Response Variable	event
Number of Response Levels	2
Weight Variable	_weight_
Model	binary logit
Optimization Technique	Fisher's scoring

Parameter	DF	Estimate
Intercept	1	-1.5533
baseage_1	1	-0.6195
baseage_2	1	-0.5315
baseage_3	1	-0.2737
smkhx	1	0.0519
treat_1	1	-0.6531
treat_2	1	-0.1870
stage_1	1	-0.0711
psa_1	1	0.1947
gleason_1	1	-0.8889
gleason_2	1	-0.3619
bmi_pre_1	1	-0.4894
bmi_pre_2	1	0.1708
bmi_pre_3	1	0.5384
vigact_pre_1	1	0.1136
vigact_pre_2	1	-0.0524
vigact_pre_3	1	0.3216
modact_pre_1	1	-0.2790
modact_pre_2	1	0.0895
modact_pre_3	1	-0.3047
fhxmi	1	-0.4512
period_1	1	-1.9668
period_2	1	-1.0941
period_3	1	-0.7179
period_4	1	-0.6624
xcond	1	1.3141
tsxcond_inter	1	-0.1149
modact	1	-0.2250
modact_spl1	1	0.1548
bmi_1	1	1.8081
bmi_2	1	0.7712
bmi_3	1	0.2690
vigact	1	-0.2727
vigact_spl1	1	0.1708

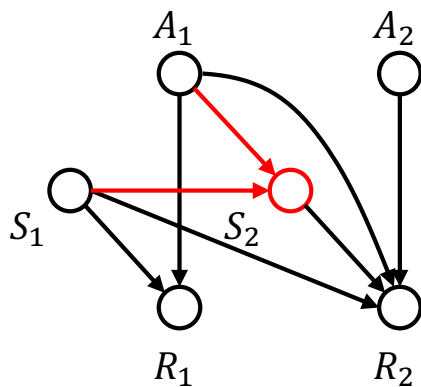


(Dickerman et al., Guideline-Based Physical Activity and Survival Among US Men With Nonmetastatic Prostate Cancer. Am J Epidemiol. 2019)

Physical activity for men with prostate cancer

Model 2 Development of conditions limiting physical activity model (composite of functional impairment, metastasis, myocardial infarction, stroke, congestive heart failure, or amyotrophic lateral sclerosis)

The LOGISTIC Procedure		Parameter	DF	Estimate
Model Information		Intercept	1	-0.0539
Data Set	WORK.PARAM	baseage_1	1	-1.2596
Response Variable	xcond	baseage_2	1	-0.5674
Number of Response Levels	2	baseage_3	1	-0.3398
		smkhx	1	0.0342
		treat_1	1	-0.5107
		treat_2	1	-0.1951
		stage_1	1	-0.2366
		psa_1	1	-0.3521
		gleason_1	1	-0.6022
		gleason_2	1	-0.3454
		bmi_pre_1	1	-1.5383
		bmi_pre_2	1	-0.0499
		bmi_pre_3	1	-0.1727
		vigact_pre_1	1	-0.0748
		vigact_pre_2	1	-0.00800
		vigact_pre_3	1	0.1594
		modact_pre_1	1	0.1080
		modact_pre_2	1	0.2414
		modact_pre_3	1	0.0816
		fhxmi	1	0.2013
		period_1	0	0
		period_2	1	-0.3085
		period_3	1	-0.3899
		period_4	1	-0.3082
		modact_l1	1	-0.0839
		modact_l1_sp11	1	0.0492
		bmi_l1_1	1	0.6708
		bmi_l1_2	1	-0.6886
		bmi_l1_3	1	-0.3326
		vigact_l1	1	-0.1384
		vigact_l1_sp11	1	0.0617



(Dickerman et al., Guideline-Based Physical Activity and Survival Among US Men With Nonmetastatic Prostate Cancer. Am J Epidemiol. 2019)

Physical activity for men with prostate cancer

Model 4 BMI model

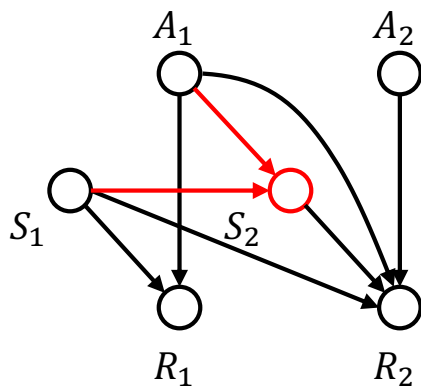
The REG Procedure
 Model: MODEL1
 Dependent Variable: bmi

Number of Observations Read	6820
Number of Observations Used	6820

Root MSE	0.06711	R-Square	0.7217
Dependent Mean	3.24274	Adj R-Sq	0.7203
Coeff Var	2.06948		

Parameter Estimates

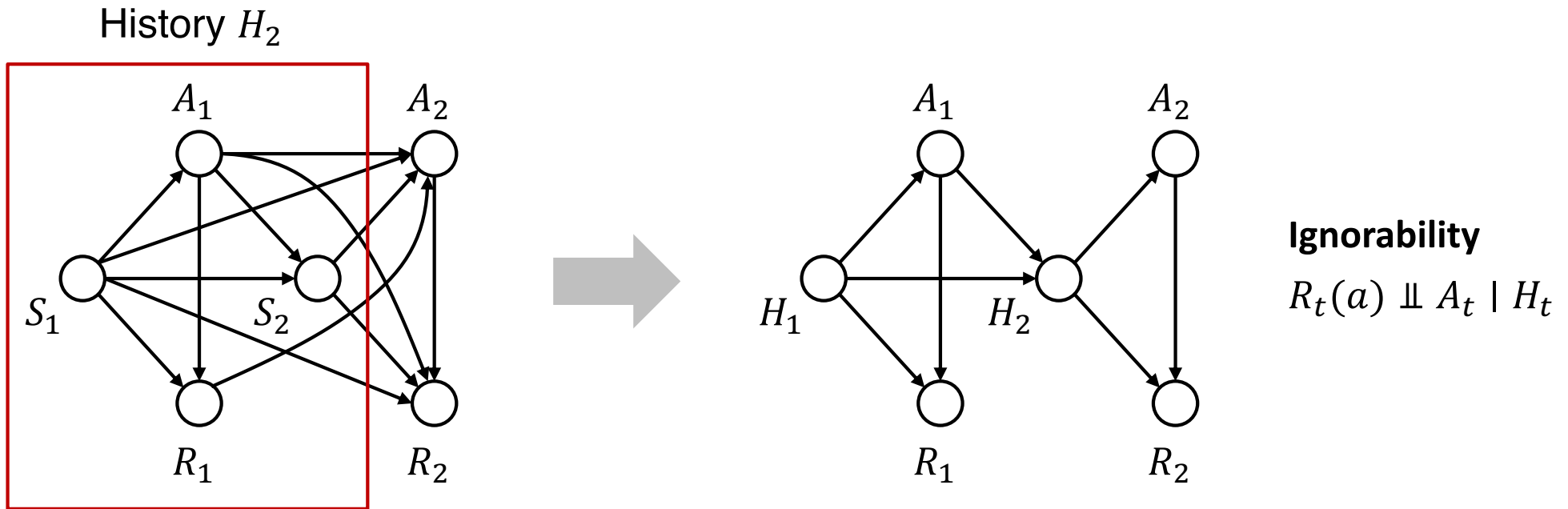
Variable	DF	Parameter Estimate	Standard Error
Intercept	1	3.46142	0.00687
baseage_1	1	0.01699	0.00271
baseage_2	1	0.01693	0.00263
baseage_3	1	0.00810	0.00249
smkhx	1	0.00309	0.00166
treat_1	1	0.00411	0.00315
treat_2	1	0.00132	0.00307
stage_1	1	-0.00279	0.00187
psa_1	1	-0.00291	0.00243
gleason_1	1	-0.00258	0.00317
gleason_2	1	-0.00532	0.00339
bmi_pre_1	1	-0.13960	0.01377
bmi_pre_2	1	-0.11586	0.00451
bmi_pre_3	1	-0.05476	0.00390
vigact_pre_1	1	0.00198	0.00273
vigact_pre_2	1	0.00270	0.00344
vigact_pre_3	1	0.00167	0.00305
modact_pre_1	1	0.00249	0.00239
modact_pre_2	1	0.00261	0.00247
modact_pre_3	1	0.00085917	0.00251
fhxmi	1	-0.00353	0.00248
period_1	0	0	.
period_2	1	0.00642	0.00255
period_3	1	0.00291	0.00258
period_4	1	0.00144	0.00266
xcond	1	-0.00474	0.00521
tsxcond_inter	1	0.00171	0.00252
modact_l1	1	0.00020698	0.00064776
modact_l1_sp11	1	-0.00006509	0.00047249
bmi_l1_1	1	-0.38170	0.01398
bmi_l1_2	1	-0.22560	0.00431
bmi_l1_3	1	-0.12555	0.00370
vigact_l1	1	-0.00171	0.00105
vigact_l1_sp11	1	0.00096989	0.00090280
modact	1	-0.00045448	0.00062425
modact_sp11	1	0.00000458	0.00045730



(Dickerman et al., Guideline-Based Physical Activity and Survival Among US Men With Nonmetastatic Prostate Cancer. Am J Epidemiol. 2019)

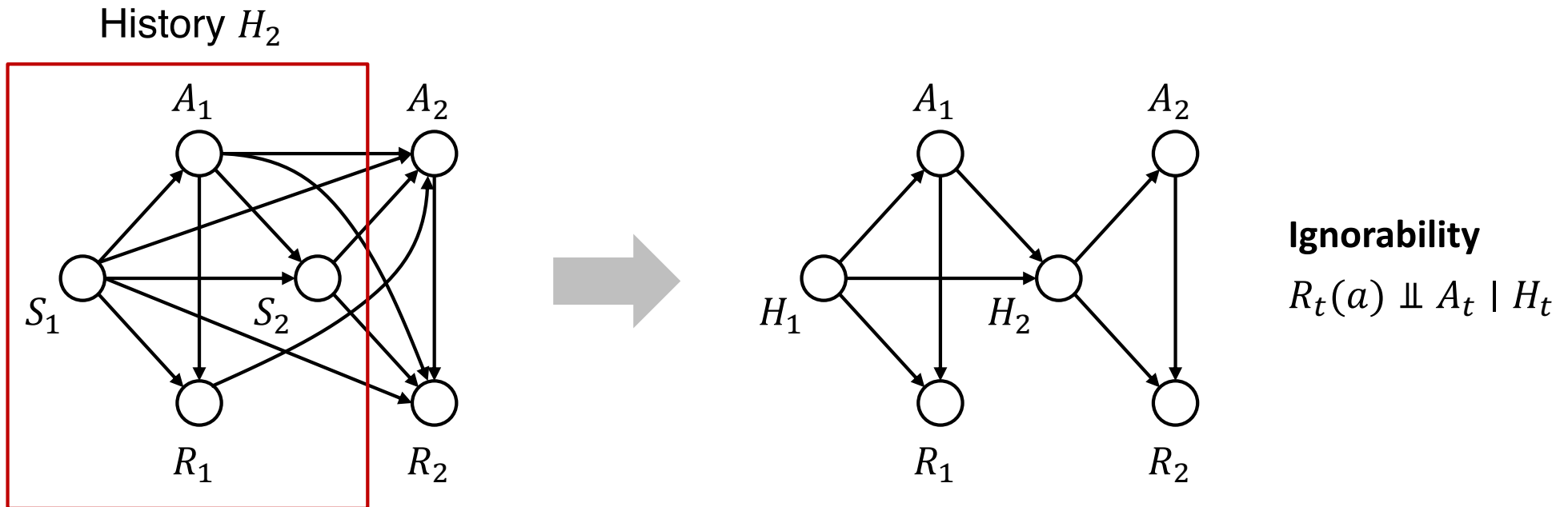
Sharing parameters for policies, time-dependent confounders, and outcomes

- To have sequential ignorability, we need to remember history



Sharing parameters for policies, time-dependent confounders, and outcomes

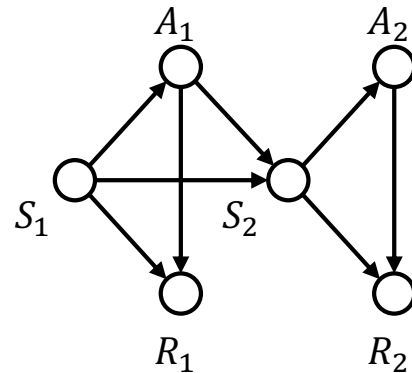
- To have sequential ignorability, we need to remember history



- The difficulty with history is that its **size grows with time**
- Use domain knowledge to summarize salient parts of history into a fixed set of time-dependent confounders
- Alternatively, **learn a summary** function that maintains what is relevant, e.g., using an RNN

Sharing parameters for policies, time-dependent confounders, and outcomes

- Look familiar? This is a Markov decision process (MDP), and we are doing (batch) reinforcement learning!



Will use S instead of H , but remember how we got here

- AlphaStar
- AlphaGo
- DQN Atari
- Open AI Five

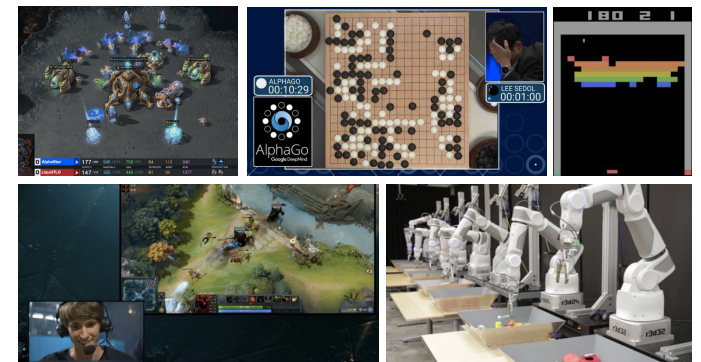
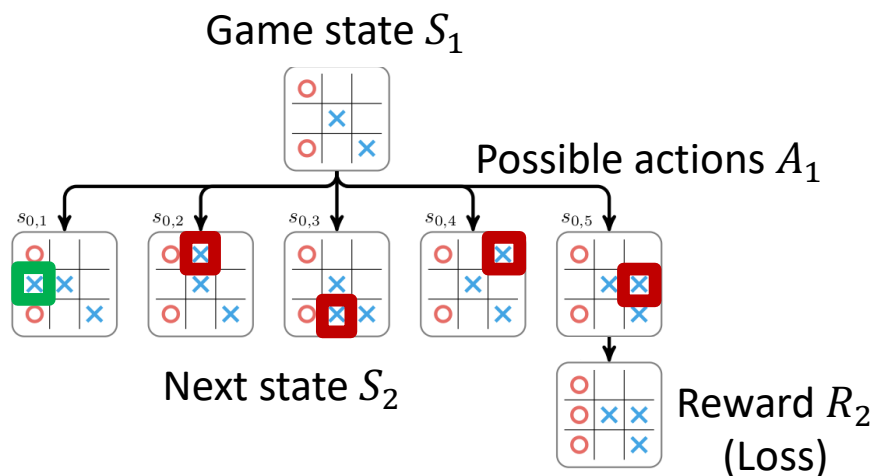


Figure by Tim Wheeler, tim.hibal.org

Sharing parameters for policies, time-dependent confounders, and outcomes

- Look familiar? This is a Markov decision process (MDP), and we are doing (batch) reinforcement learning!
- Up until now, we have only talked about *evaluation* of dynamic treatment regimes
- How do we find **optimal** policies?
 1. Policy gradient using G-computation (estimate MDP first) or marginal structural models (inverse propensity score-based estimator)
 2. Dynamic programming (G-estimation) or Q-learning

Summary

Significant care needed when performing off-policy RL in healthcare

- What are the decision points?
- What is the underlying causal graph? (Taking into consideration clinical practice today.)
- Is there hidden confounding? When does positivity (overlap) hold?
- What are reasonable ways to share parameters without *creating* hidden confounding?

Consider tackling evaluation of a few reasonable policies before attempting to use black-box methods to learn an optimal policy

Additional references

- Chakraborty & Moodie, [Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine](#). Springer, 2013
- O. Gottesman, F. Johansson, M. Komorowski, A. Faisal, D. Sontag, F. Doshi-Velez, L. Celi. [Guidelines for reinforcement learning in healthcare. Nature medicine](#), 2019
- Li et al., [G-Net: a Recurrent Network Approach to G-Computation for Counterfactual Prediction Under a Dynamic Treatment Regime](#). Proceedings of Machine Learning Research 158:282–297, 2021
- Hua, Mei, Zohar, Giral, Xu. [Personalized Dynamic Treatment Regimes in Continuous Time: A Bayesian Approach for Optimizing Clinical Decisions with Timing](#). Bayesian Analysis. Advance Publication, 1-30, 2021