Personalizing renal replacement therapy initiation in the intensive care unit: a statistical reinforcement learning-based strategy with external validation on the AKIKI randomized controlled trials

François Grolleau
Chef de Clinique Assistant — Hôpital Hôtel Dieu (APHP)

AI for Science, Science for AI — CNRS
Causality in Practice Seminar

Wednesday 14, June 2023

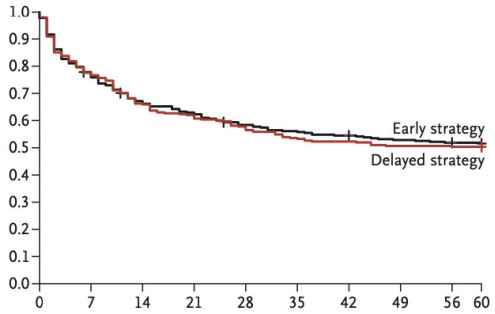
Université Paris Cité

CRESS Epidemiology Statistics

What is Renal-Replacement Therapy (RRT)?

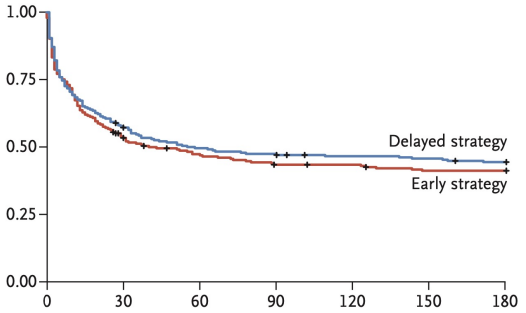
Randomized controlled trials on RRT timing

AKIKI. NEJM. 2016.



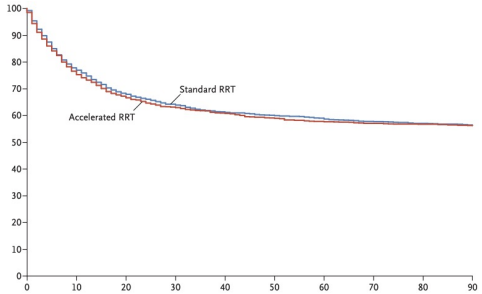
HR 0.97 [0.77 – 1.21]

IDEAL-ICU. NEJM. 2018.



HR 1.08 [0.84 – 1.38]

STARRT-AKI. NEJM. 2020.



RR 1.00 [0.93 – 1.09]
90 days

AKIKI-2. Lancet. 2021.

"60-day mortality did not differ significantly between groups."

A request from the medical community

Comment of the AKIKI 2 trial in the Lancet[1]

"Results of the AKIKI 2 trial and other timing studies should motivate us to complement clinical judgment with new strategies, including **dynamic** […] **decision tools** to enable personalised management."

[1]Wait and see for acute dialysis: but for how long? M. Ostermann et al. Lancet. 2021.

Static decision support tool

Plug values from the time point when severe acute kidney injury occurs* (KDIGO III or RIFLE failure stage).

SOFA
range in training data: 3 to 21
10

pH
range in training data: 6.88 to 7.54
7.3

Potassium (mmol/L)
range in training data: 2.4 to 7.4 mmol/L
4.5

Urea (mmol/L)
range in training data: 2 to 59 mmol/L
20

Weight (kg)
range in training data: 34 to 200 kg
80

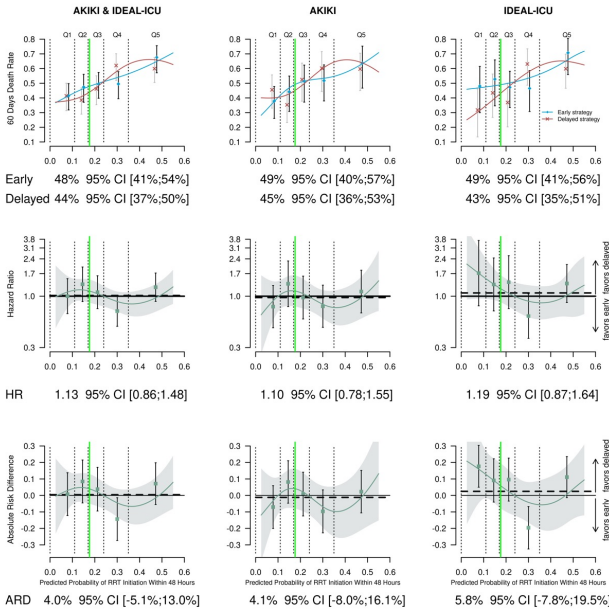
Immunosuppressive Drug (non-corticosteroid)
- Yes
- No

Predict

**Provided the patient meets inclusion/exclusion criterion for the AKIKI or IDEAL-ICU trials.*

http://rrt-personalization.eu/

The predicted probablity of RRT initiation within 48 hours is 18% which corresponds to the 60-day mortality outcome below.



	AKIKI & IDEAL-ICU	AKIKI	IDEAL-ICU
Early	48% 95% CI [41%;54%]	49% 95% CI [40%;57%]	49% 95% CI [41%;56%]
Delayed	44% 95% CI [37%;50%]	45% 95% CI [36%;53%]	43% 95% CI [35%;51%]
HR	1.13 95% CI [0.86;1.48]	1.10 95% CI [0.78;1.55]	1.19 95% CI [0.87;1.64]
ARD	4.0% 95% CI [-5.1%;13.0%]	4.1% 95% CI [-8.0%;16.1%]	5.8% 95% CI [-7.8%;19.5%

RRT = Renal-Replacement Therapy; CI = Confidence Interval; HR = Hazard Ratio; ARD = Absolute Risk Difference

Using data from AKIKI & IDEAL-ICU, this tool makes personalized recommendation <u>at a single point in time</u>:
t_0: acute kidney injury (AKI) stage 3 KDIGO

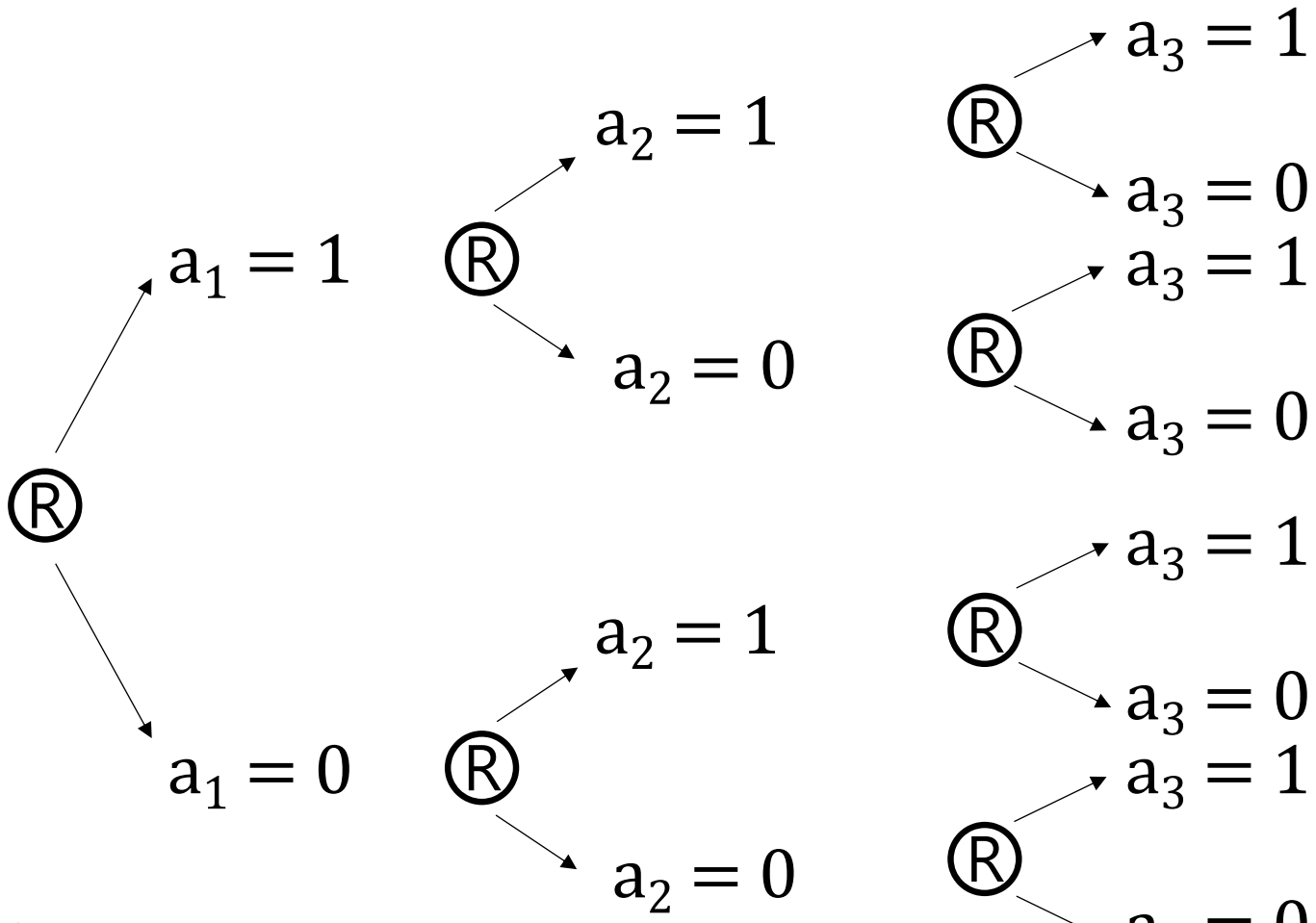
Start RRT now

or

Wait and reevaluate (the next day)

F. Grolleau, R.Porcher et al. *Critical Care* 2022.

What is the target trial we aim to “emulate”?



® randomization
a_t = 1 : start RRT in the 24 hours following decision point *t*
a_t = 0 : do not start RRT in the 24 hours following decision point *t*

[1]Sequential Multiple Assignment Randomized Trial (Almirall D et al., Transl Behav Med 2014)

Study objective

1. Estimate an optimal dynamic strategy for RRT initiation in the ICU by using reinforcement learning methods on EHR data (development step)

2. Evaluate this strategy on data from two RCTs (external validation step)

Questions?

Sources of data

1. Development: MIMIC-III

Observational EHR data from 61 051 ICU adult admissions at a tertiary hospital in Boston (BIDMC — Harvard Medical School) between 2001 and 2012.

2. Validation: AKIKI and AKIKI2 trials

Compared (early *vs* delayed) and (delayed *vs* more delayed) strategies respectively

- We only use data from the delayed arm which are now considered the “best practices”
- We analyzed these data as “high quality” observational data—not exploiting the randomization design

Eligibility criteria

Identical to the eligibility criteria from the original AKIKI trial

- Inclusion

ICU adult patients with stage 3 KDIGO acute kidney injury
Who required mechanical ventilation, catecholamine infusion, or both

- Exclusion

Patient in a moribund state
Patient with end-stage kidney disease
Patients already included at a previous date

Problem formalization

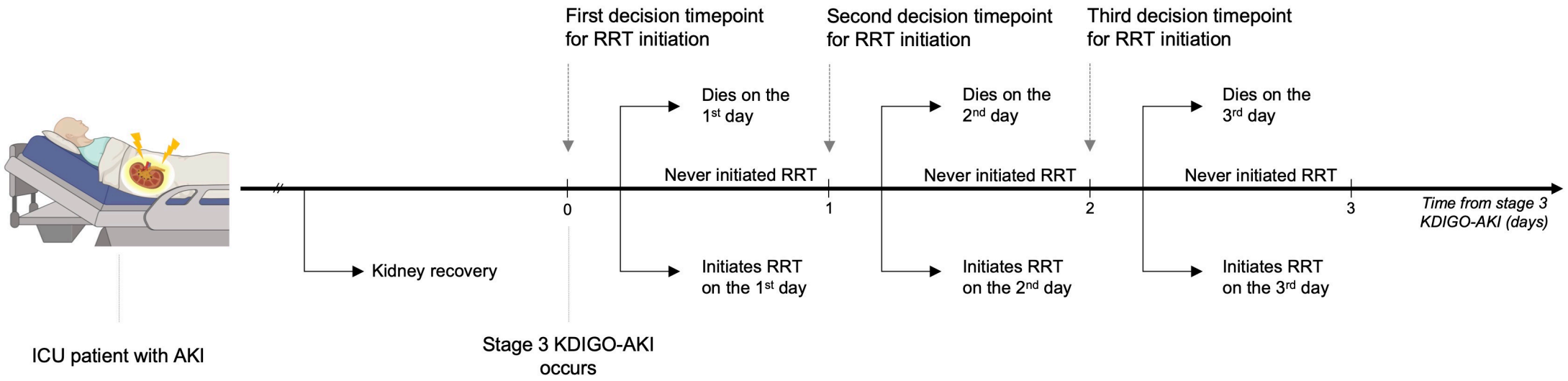


Figure 1 Possible trajectories of a single patient with acute kidney injury in our learning setup.

Notations

$t \in \{1,2,3\}$	Decision point
S_t	Patient state (features) just before decision t is taken
$A_t \in \{0,1\}$	Action taken at decision t $A_t = 1$: on RRT in the 24 hours following decision point t $A_t = 0$: off RRT in the 24 hours following decision point t
$H_t \coloneqq (S_1, A_1, \ldots, S_t)$	Patient history up to decision t
$e_t(H_t) \coloneqq \mathbb{E}[A_t \mid H_t]$	Relevant generalization of the propensity score
$\pi \coloneqq \{\pi_t\}_{t=1}^{3}$ where $\pi_t : \mathcal{H}_t \mapsto \{0,1\}$	A deterministic, nonstationary strategy

Distal reward to maximize

Y: Hospital-free days at day 60

1. Patient relevant: “number of days alive and outside the hospital at day 60”

2. Prevent “value misalignment”[3] which could occur in case of maximization of 60-day survival. e.g., an optimal strategy maximizes 60-day survival at the expense of higher hospital mortality or ICU mortality.

[3]Sutton RS, Barto AG. 17.4 Designing reward signals. *Reinforcement learning: An introduction* MIT press; 2018.

Questions?

Our objective

Within a strategy class Π, we wish to estimate an optimal strategy

$$\pi^{opt} := \arg\max_{\pi \in \Pi} \mathbb{E}\left[Y^{\pi_1(H_1),\pi_2(H_2),\pi_3(H_3)}\right].$$

That is, a strategy that if followed, would maximize the expected reward.

Importantly, we need to explicitly or implicitly decide what the strategy class Π is
If its too high dimensional → severe risk of overfitting

Problem recap and contrast with traditional RL

$\overline{a}_t = a_1, \ldots, a_t$

$\underline{a}_t = a_t, \ldots, a_T$

Observational data → Offline (i.e., batch) RL

Finite horizon for decision i.e., $T = 3$ → No need to discount rewards as in typical MDP/POMDPs

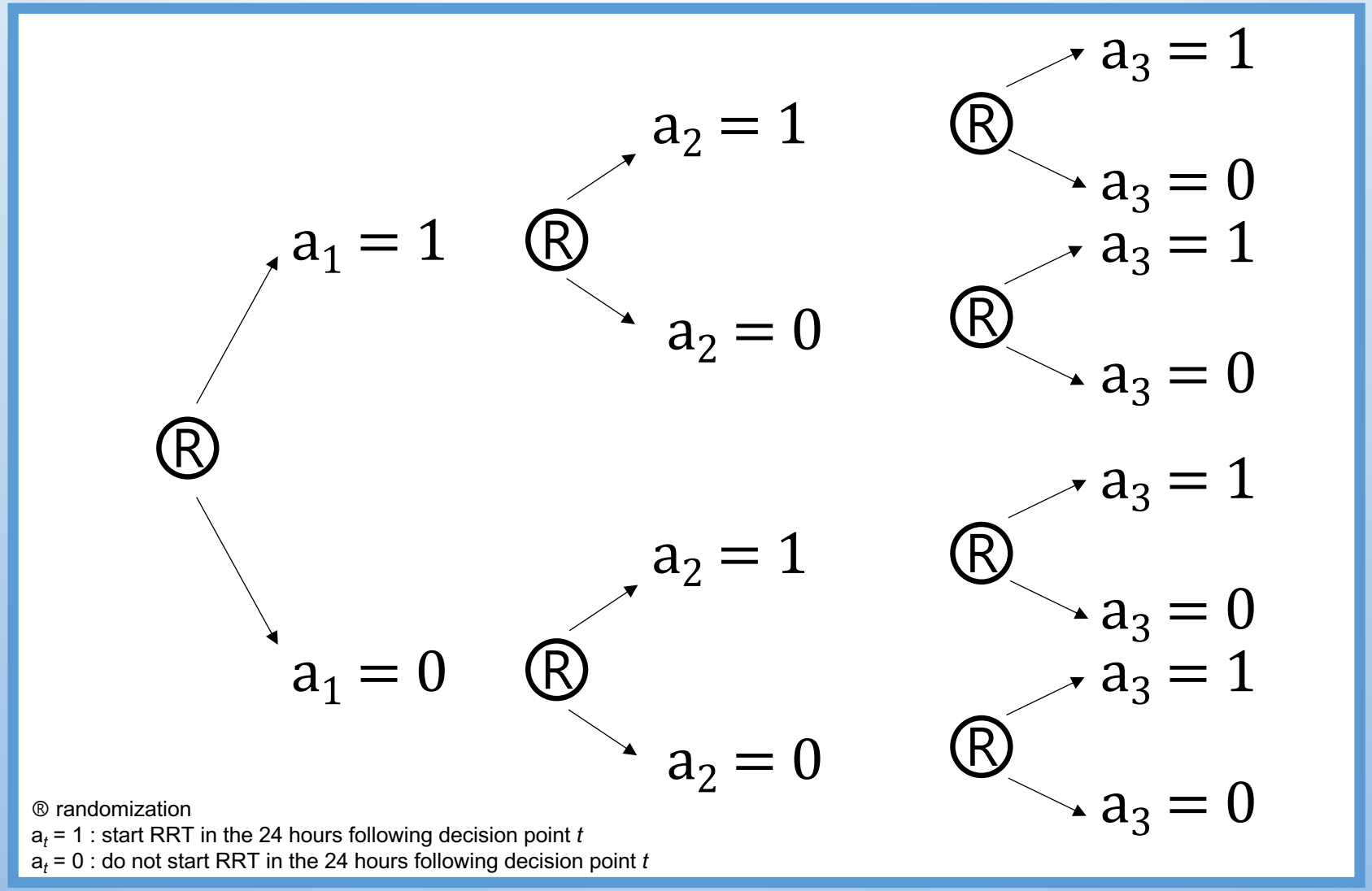
Single distal reward → Sparse reward problem!

Markov assumption are often unreasonable in medicine → Sequential Randomization assumption i.e., for all $\underline{a}_t$,

$$\left\{ (S_{t'}^{\overline{A}_{t-1}, a_t, \ldots, a_{t'}})_{t'=t+1}^{T}, Y^{\overline{A}_{t-1}, \underline{a}_t} \right\} \perp\!\!\!\perp A_t | H_t, \quad \forall t = 1, \ldots, T.$$

Our problem fits in the dynamic treatment regime (DTR) setting—which is a subset of RL.
DTR methods tackle harder problems than typical online RL algorithms that rely on MDP/POMDPs

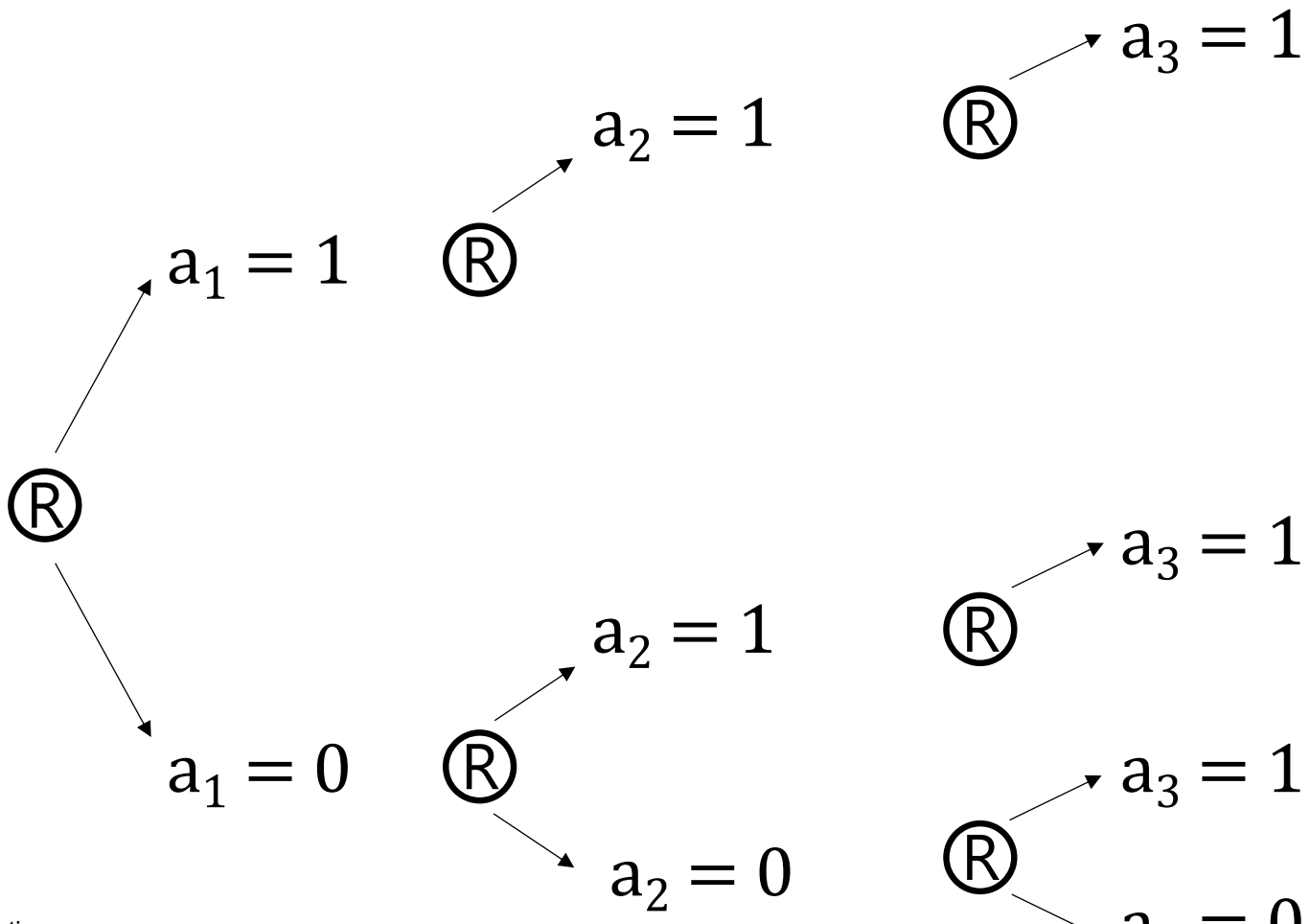
We are not interested in stopping RRT once it has been initiated



® randomization
a_t = 1 : start RRT in the 24 hours following decision point *t*
a_t = 0 : do not start RRT in the 24 hours following decision point *t*

[1]Sequential Multiple Assignment Randomized Trial (Almirall D et al., Transl Behav Med 2014)

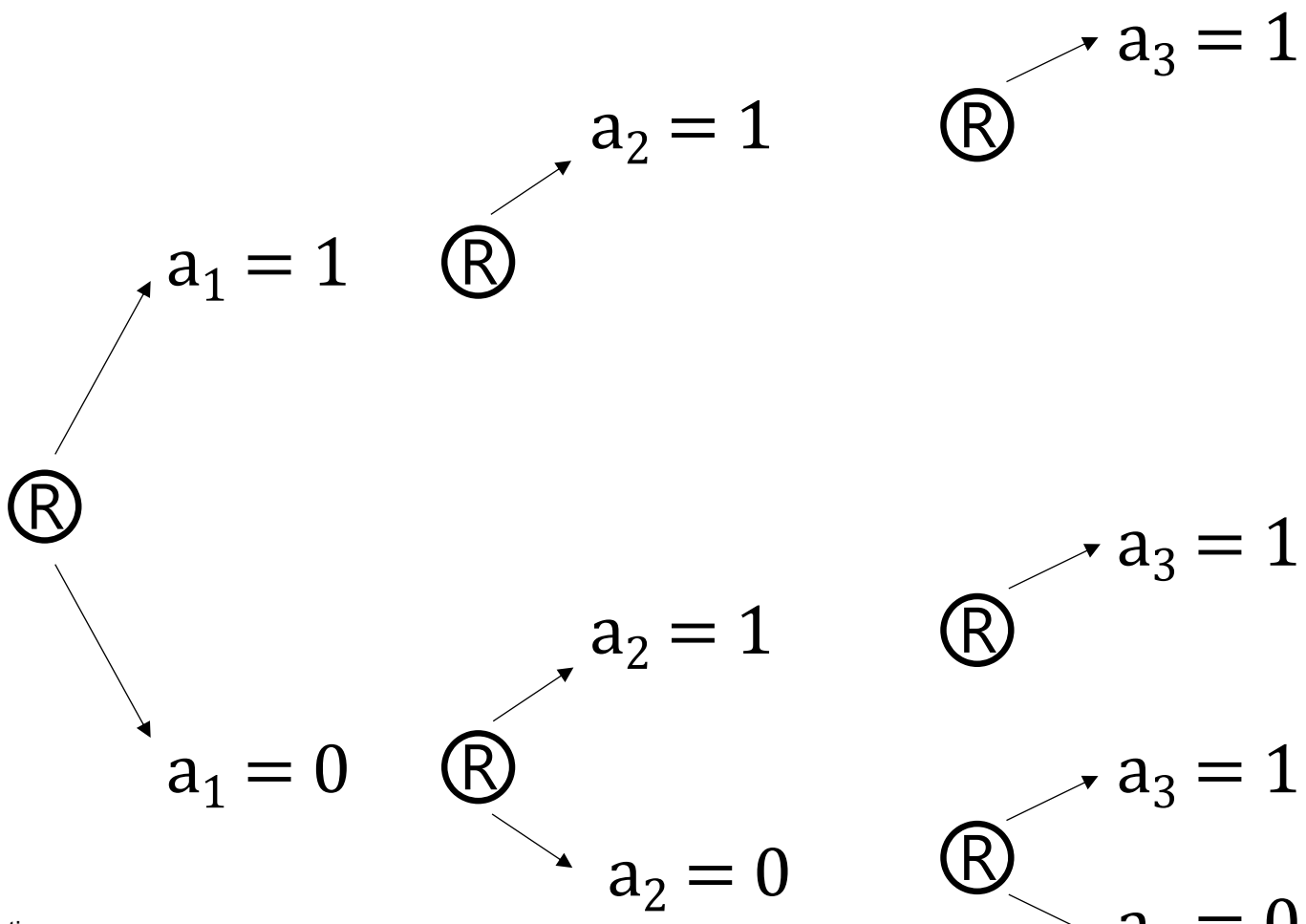
We are not interested in stopping RRT once it has been initiated



® randomization
a_t = 1 : start RRT in the 24 hours following decision point *t*
a_t = 0 : do not start RRT in the 24 hours following decision point *t*

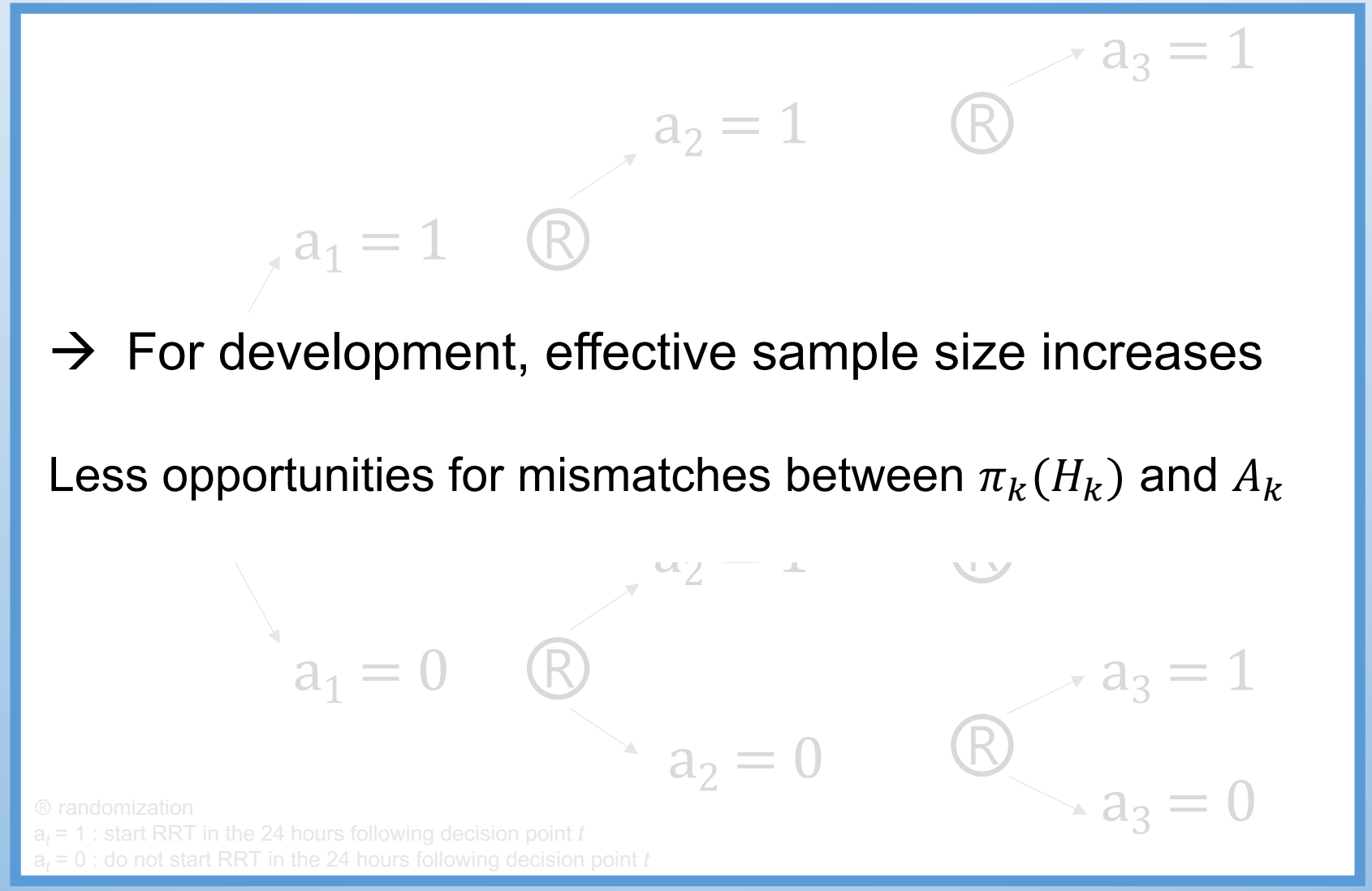
[1]Sequential Multiple Assignment Randomized Trial (Almirall D et al., Transl Behav Med 2014)

Dimensionality of the action space drops from 8 to 4



[1]Sequential Multiple Assignment Randomized Trial (Almirall D et al., Transl Behav Med 2014)

Dimensionality of the action space drops from 8 to 4



Practical consideration

How should we deal with the patients who died before the last decision point?

- Option A: Account for them explicitly e.g., define a terminal state Φ and use
$$e_t(\Phi) = 0 \text{ and } \pi_t(\Phi) = 0$$

- Option B: Exclude them from the development sample!
  - Would induce selection bias
  - Worse thing that can happen is estimate a suboptimal strategy
  - This is an acceptable so long we don't exclude patients from the validation sample

For clinical reason, we deliberately chose B.
"It seemed unlikely that the patients who died within 3 days would be released from the ICU, and subsequently from the hospital under a different RRT initiation strategy."

Estimate an optimal strategy via blip functions[3]

Consider the "blip functions" $\gamma_1(\cdot), \gamma_2(\cdot), \gamma_3(\cdot)$ such that

$$\gamma_t(a_t, h_t) := \mathbb{E}\left[Y^{\bar{a}_{t-1}, a_t, \underline{a}^{opt}_{t+1}} - Y^{\bar{a}_{t-1}, 0, \underline{a}^{opt}_{t+1}} \middle| H_t = h_t\right].$$

Given our setup, the "blip" $\gamma_t(1, h_t)$ has interpretation, the individualized treatment effect of :

Start treatment at decision point t

vs

Do not start treatment at t but deliver optimal treatment from decision point $t + 1$ onward

[3]Wallace et al. Doubly robust dynamic treatment regimen via weighted least squares. Biometrics. 2015.

Doubly robust dynamic treatment regimen via weighted least squares (dWOLS)

For $t = \{3,2,1\}$, estimate the propensity scores, $e_t(H_t) \coloneqq \mathbb{E}[A_t|H_t]$

At $t = 3$, estimate $\hat{\gamma}_3$ directly
by regressing Y onto $(H_3^f\ ,\ A_3H_3^\gamma)$ via weighted least squares with weights $\widetilde{w}_3(H_3) = |A_3\ -\ \hat{e}_3(H_3)|$

$$\rightarrow \hat{\gamma}_3\ (a_3, h_3) = \gamma_3\left(a_3, h_3;\ \hat{\beta}_3\right) = \hat{\beta}_3^T a_3 H_3^\gamma$$

Regret from decision 3

At $t = 2$,
Build the pseudo-outcome $\tilde{Y}_2 = Y + \max\{\hat{\gamma}_3(1, H_t)\ ,\ 0\} - \hat{\gamma}_3(A_3, H_3)$
Regress $\tilde{Y}_2$ onto $(H_2^f\ ,\ A_2H_2^\gamma)$ via weighted least squares with weights $\widetilde{w}_2(H_2) = |A_2\ -\ \hat{e}_2(H_2)|$

$$\rightarrow \hat{\gamma}_2\ (a_2, h_2) = \gamma_2\left(a_2, h_2;\ \hat{\beta}_2\right) = \hat{\beta}_2^T a_2 H_2^\gamma$$

A $t = 1$,
Build the pseudo-outcome $\tilde{Y}_1 = Y +$ Regret from decision 3 $+$ Regret from decision 2
Regress $\tilde{Y}_1$ onto $(H_1^f\ ,\ A_1H_1^\gamma)$ via weighted least squares with weights $\widetilde{w}_1(H_1) = |A_1\ -\ \hat{e}_1(H_1)|$

$$\rightarrow \hat{\gamma}_1\ (a_1, h_1) = \gamma_1\left(a_1, h_1;\ \hat{\beta}_1\right) = \hat{\beta}_1^T a_1 H_1^\gamma$$

[3]Wallace et al. Doubly robust dynamic treatment regimen via weighted least squares. Biometrics. 2015.

Doubly robust dynamic treatment regimen via weighted least squares (dWOLS)

For $t = \{3,2,1\}$, estimate the propensity scores, $e_t(H_t) \coloneqq \mathbb{E}[A_t|H_t]$

At $t = 3$, estimate $\hat{\gamma}_3$ directly
by regressing Y onto $(H_3^f\ ,\ A_3 H_3^\gamma)$ via weighted least squares with weights $\widetilde{w}_3(H_3) = |A_3 - \hat{e}_3(H_3)|$

Overlapping weighs of form $\widetilde{w}_t(H_t) = |A_t - \hat{e}_t(H_t)|$
Provide double robustness and enhance sample efficiency

A $t = 1$,
Build the pseudo-outcome $\tilde{Y}_1 = Y +$ Regret from decision 3 $+$ Regret from decision 2
Regress $\tilde{Y}_1$ onto $(H_1^f\ ,\ A_1 H_1^\gamma)$ via weighted least squares with weights $\widetilde{w}_1(H_1) = |A_1 - \hat{e}_1(H_1)|$

$$\rightarrow \hat{\gamma}_1\ (a_1, h_1) = \gamma_1\left(a_1, h_1;\ \hat{\beta}_1\right) = \hat{\beta}_1^T a_1 H_1^\gamma$$

[3]Wallace et al. Doubly robust dynamic treatment regimen via weighted least squares. Biometrics. 2015.

Questions?

External validation of the estimated optimal strategy

Advantage double robust estimator[4] (with terminal states) for

$$\Delta(\pi, \mathbf{0}) := \mathbb{E}[Y^{\pi_1(H_1),\ldots,\pi_T(H_T)} - Y^{0,\ldots,0}]$$
$$= \mathbb{E}_\pi[Y] - \mathbb{E}_0[Y]$$

- Double robust, sample-efficient estimator—compared to other methods
- Accounts for the patients who died before the last decision point
- Allows to compare an estimated strategy π to any other strategy e.g., a best practice strategy π_{bp}

$$\hat{\Delta}(\pi, \pi_{bp}) = \hat{\Delta}(\pi, \mathbf{0}) - \hat{\Delta}(\pi_{bp}, \mathbf{0})$$

[4]Nie et al. Learning when-to-treat policies. JASA. 2021.

Anatomy of the ADR estimator with terminal states

$$\hat{\Delta}(\pi, \mathbf{0}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbb{1}_{S_t \neq \Phi} \mathbb{1}_{t \geq \tau_\pi^{(i)}} \frac{\mathbb{1}_{A_{1:t-1}^{(i)}=0}}{\prod_{t'=1}^{t-1} \hat{e}_{t',0}^{-q(i)}(S_{1:t'}^{(i)})} \hat{\Psi}_{t,W_\pi}^{\Phi}(S_{1:t}^{(i)})$$

$S_t = \Phi$ for dead patients

Product of propensity scores

Sum over all decision points

where the relevant DR score is

$$\hat{\Psi}_{t,k}^{\Phi}(S_{1:t}^{(i)}) = \hat{\mu}_{\text{now},k}^{-q(i)}(S_{1:t}^{(i)}, t) - \hat{\mu}_{\text{next},k}^{\Phi}(S_{1:t}^{(i)}, t)^{-q(i)}$$
$$+ \mathbb{1}_{A_t^{(i)}=k} \frac{Y^{(i)} - \hat{\mu}_{\text{now},k}^{-q(i)}(S_{1:t}^{(i)}, t)}{\hat{e}_{t,k}^{-q(i)}(S_{1:t}^{(i)})}$$
$$- \mathbb{1}_{A_t^{(i)}=0} \mathbb{1}_{A_{t+1}^{(i)}=k} \frac{Y^{(i)} - \hat{U}^{-q(i)}(S_{1:t}^{(i)}, \Phi)}{\hat{e}_{t,0}^{-q(i)}(S_{1:t}^{(i)}) \hat{e}_{t+1,k}^{-q(i)}(S_{1:t+1}^{(i)})},$$

[4]Nie et al. Learning when-to-treat policies. JASA. 2021.

Anatomy of the ADR estimator with terminal states

$$\hat{\Delta}(\pi, \mathbf{0}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbb{1}_{S_t \neq \Phi} \mathbb{1}_{t \geq \tau_\pi^{(i)}} \frac{\mathbb{1}_{A_{1:t-1}^{(i)}=0}}{\prod_{t'=1}^{t-1} \hat{e}_{t',0}^{-q(i)}(S_{1:t'}^{(i)})} \hat{\Psi}_{t,W_\pi}^{\Phi}(S_{1:t}^{(i)})$$

← Only part that is double robust, i.e., the estimator itself is not double robust

where the relevant DR score is

$$\hat{\Psi}_{t,k}^{\Phi}(S_{1:t}^{(i)}) = \hat{\mu}_{\text{now},k}^{-q(i)}(S_{1:t}^{(i)}, t) - \hat{\mu}_{\text{next},k}^{\Phi}(S_{1:t}^{(i)}, t)^{-q(i)}$$
$$+ \mathbb{1}_{A_t^{(i)}=k} \frac{Y^{(i)} - \hat{\mu}_{\text{now},k}^{-q(i)}(S_{1:t}^{(i)}, t)}{\hat{e}_{t,k}^{-q(i)}(S_{1:t}^{(i)})}$$
$$- \mathbb{1}_{A_t^{(i)}=0} \mathbb{1}_{A_{t+1}^{(i)}=k} \frac{Y^{(i)} - \hat{U}^{-q(i)}(S_{1:t}^{(i)}, \Phi)}{\hat{e}_{t,0}^{-q(i)}(S_{1:t}^{(i)}) \hat{e}_{t+1,k}^{-q(i)}(S_{1:t+1}^{(i)})},$$

← Interpretation is "the relative advantage of

starting treatment at t
versus
starting at $t+1$

given an individual history up to t."

This estimator relies on a decomposition into a sum of local advantages as proposed by Murphy, JMLR 2005 (*see* Lemma 1)

[4]Nie et al. Learning when-to-treat policies. JASA. 2021.

Condition for the ADR proof

The evaluated when-to-treat strategy needs be *regular*[2] in the sense that

"If the estimated strategy π suggest to start RRT at a given decision point t, it persists in this choice even if physicians fail to start treatment immediately."

→ Formalizes the off-policy behavior of π

[2]Definition 1 in Nie, Brunskill, Wager, JASA 2020.

Anatomy of the ADR estimator with terminal states

$$\hat{\Delta}(\pi, \mathbf{0}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbb{1}_{S_t \neq \Phi} \mathbb{1}_{t \geq \tau_\pi^{(i)}} \frac{\mathbb{1}_{A_{1:t-1}^{(i)}=0}}{\prod_{t'=1}^{t-1} \hat{e}_{t',0}^{-q(i)}(S_{1:t'}^{(i)})} \hat{\Psi}_{t,W_\pi}^{\Phi}(S_{1:t}^{(i)})$$

where the relevant DR score is

$$\hat{\Psi}_{t,k}^{\Phi}(S_{1:t}^{(i)}) = \hat{\mu}_{\text{now},k}^{-q(i)}(S_{1:t}^{(i)}, t) - \hat{\mu}_{\text{next},k}^{\Phi}(S_{1:t}^{(i)}, t)^{-q(i)}$$

$$+ \mathbb{1}_{A_t^{(i)}=k} \frac{Y^{(i)} - \hat{\mu}_{\text{now},k}^{-q(i)}(S_{1:t}^{(i)}, t)}{\hat{e}_{t,k}^{-q(i)}(S_{1:t}^{(i)})}$$

$$- \mathbb{1}_{A_t^{(i)}=0} \mathbb{1}_{A_{t+1}^{(i)}=k} \frac{Y^{(i)} - \hat{U}^{-q(i)}(S_{1:t}^{(i)}, \Phi)}{\hat{e}_{t,0}^{-q(i)}(S_{1:t}^{(i)}) \hat{e}_{t+1,k}^{-q(i)}(S_{1:t+1}^{(i)})},$$

For all nuisance parameters, the authors suggest cross-fitting to reduce the effect of own-observation bias

[4]Nie et al. Learning when-to-treat policies. JASA. 2021.

Questions?

Estimated blip functions yield linear thresholding strategies

- For decision point $t = 1$ (AKI KDIGO stage 3), start RRT within 24 hours if

$$39.589 + 0.245\times\text{age}_{t=1}\ (years) + 1.349\times\text{creatinine}_{t=1}\ (mg/dL) + 3.409\times\text{potassium}_{t=1}\ (mmoL/L) > 0$$

- For decision point $t = 2$ (AKI KDIGO stage 3 + 24 hours), start RRT within 24 hours if

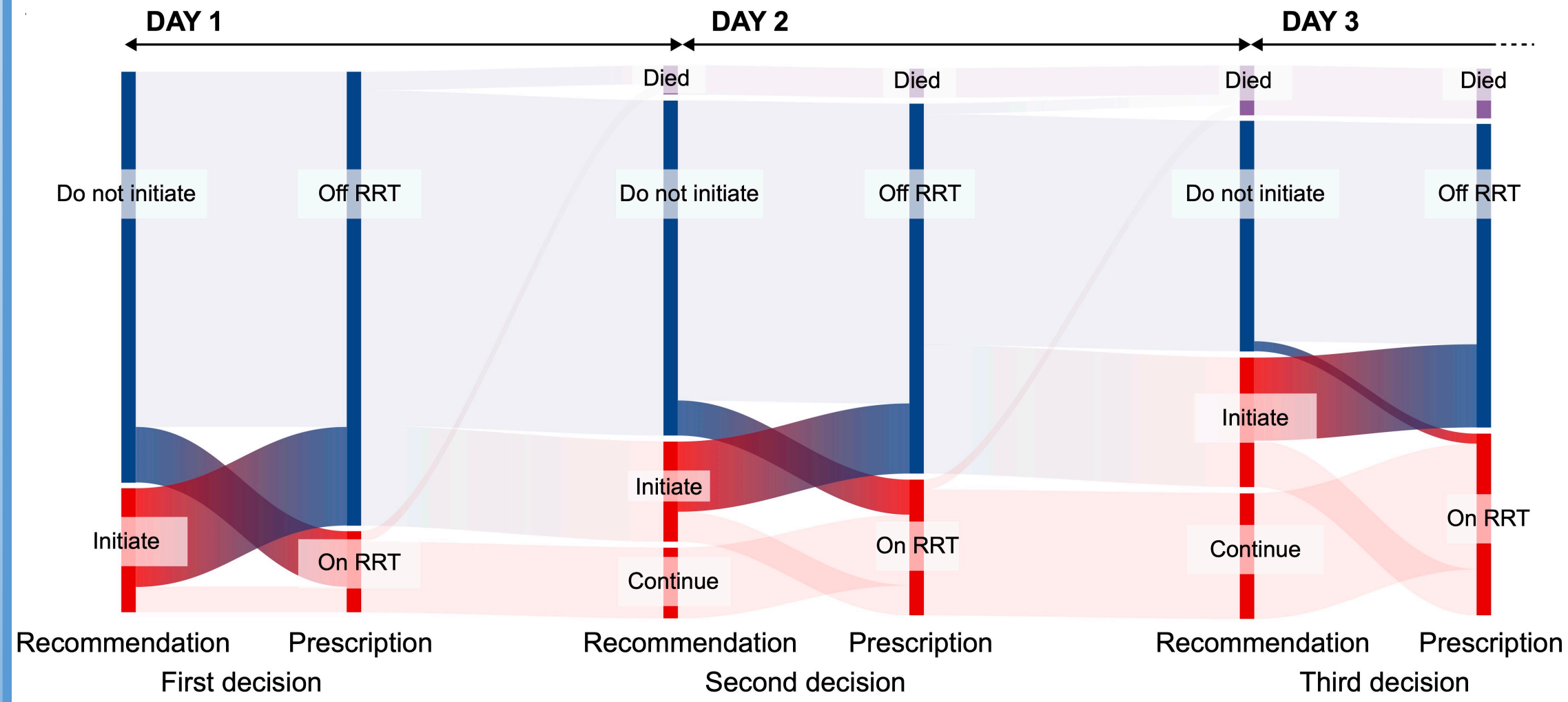
$$-7.747 + 0.514\times\text{SOFA}_{t=2} + 0.095\times\text{blood urea nitrogen}_{t=2}\ (mg/dL) - 63.874\times|\text{pH}_{t=1} - \text{pH}_{t=2}| - 7.734\times[\text{urine output}_{t=1} + \text{urine output}_{t=2}] > 0$$

- For decision point $t = 3$ (AKI KDIGO stage 3 + 48 hours), start RRT within 24 hours if

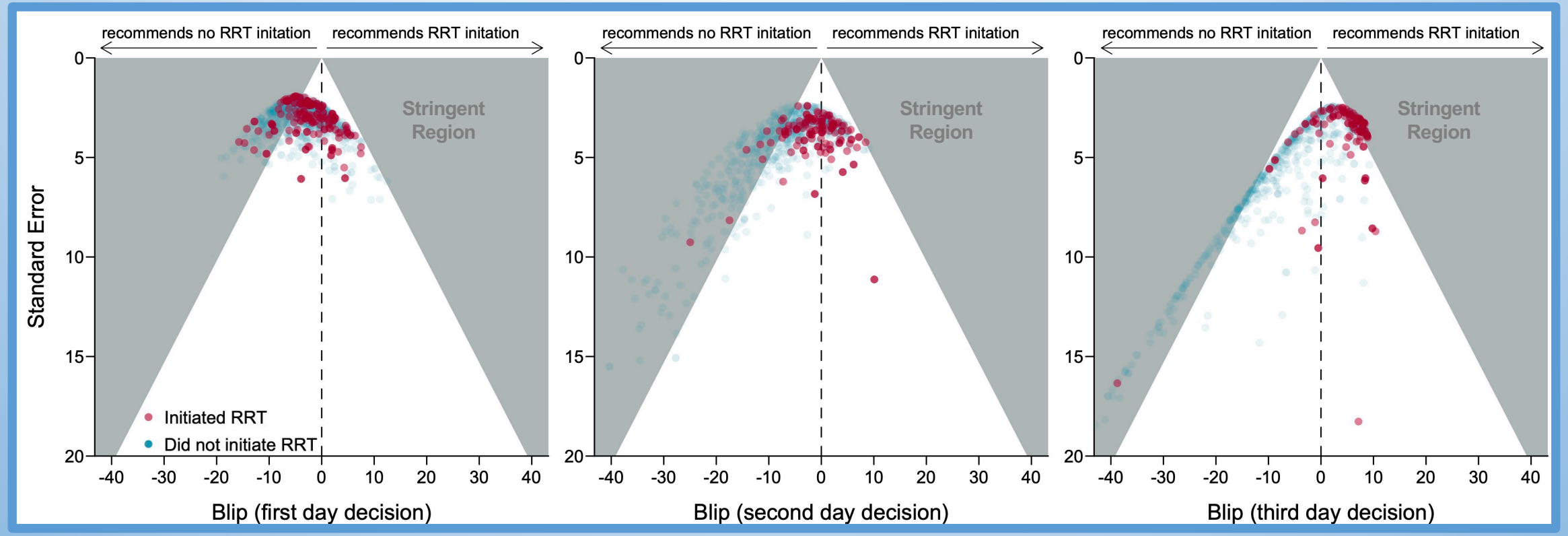
$$5.397 - 19.316\times\text{urine output}_{t=3}\ (ml/kg/h) + 1.922\times[\text{blood urea nitrogen}_{t=3} / \text{blood urea nitrogen}_{t=1}] > 0$$

→ Recall that the strategy was not built to recommend stopping RRT once it has been initiated

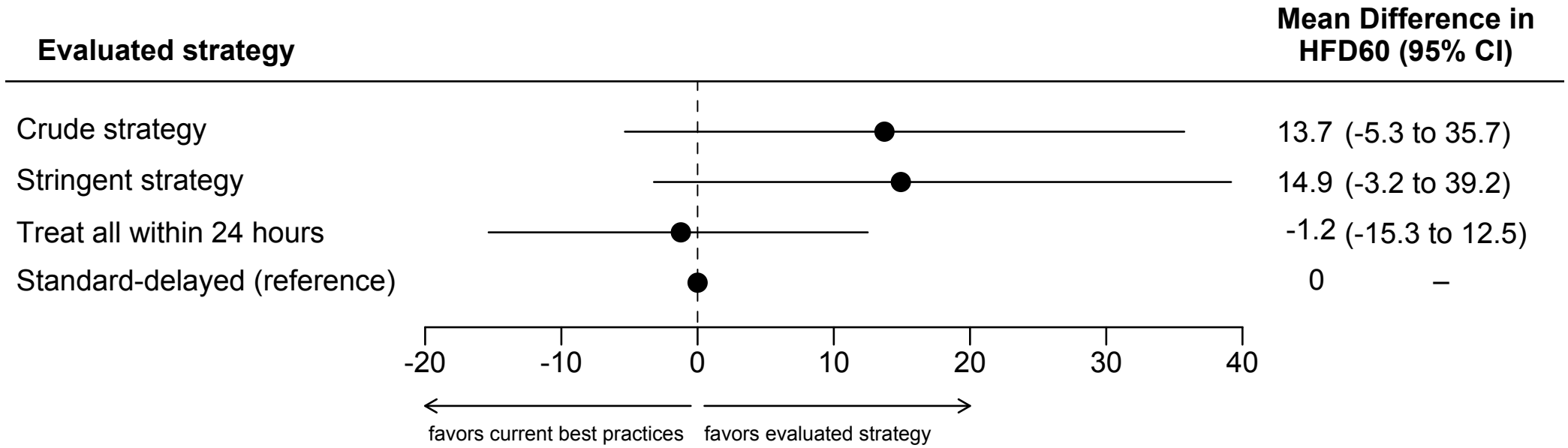
Discrepancies between current best practices and the recommendations from the learned strategy



Individual patient recommendations (validation set)

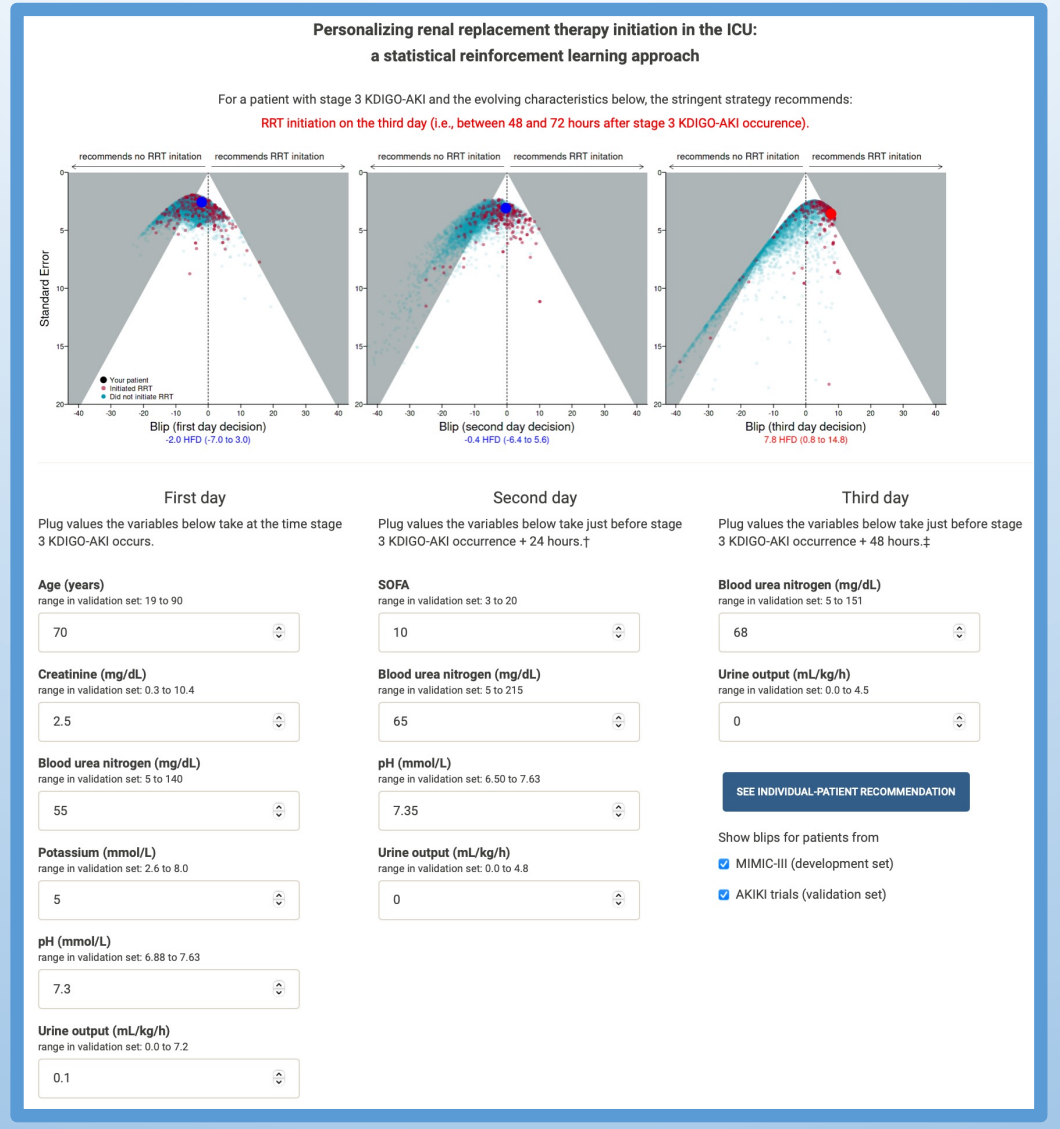


Strategy evaluation (external validation via ADR estimator)



Questions?

Computerized decision support tool: http://dynamic-rrt.eu/



Interpretation and perspectives

- We developed a practical and interpretable dynamic decision support system for RRT initiation in the ICU
- External validation shows that its implementation could improve the average number of days that ICU patients spend alive and outside the hospital
- In the next step, infrastructure work needs be done to automate the extraction of data for relevant to RRT initiation strategies. This would reduce clinical burden and allow
  i. Testing the tool in clinical settings
  ii. Reevaluate the tool on routinely collected data
  iii. Redevelop or recalibrate the tool to improve its effectiveness

Thank you!

Questions?

- We developed a practical and interpretable dynamic decision support system for RRT initiation in the ICU.
- External validation shows that its implementation could improve the average number of days that ICU patients spend alive and outside the hospital.
- In the next step, infrastructure work needs be done to automate the extraction of data for relevant to RRT initiation strategies. This would reduce clinical burden and allow
  i. Testing the tool in clinical settings
  ii. Reevaluate the tool on routinely collected data
  iii. Redevelop or recalibrate the tool to improve its effectiveness

Algorithm 1: Advantage doubly robust (ADR) estimator

1 Estimate the outcome models $\mu_{\text{now},k}(\cdot)$, $\mu_{\text{next},k}(\cdot)$, as well as treatment propensities $e_{t,a}(s_{1:t})$ with cross-fitting using any supervised learning method tuned for prediction accuracy.

2 Given these nuisance component estimates, we construct value estimates

$$\hat{\Delta}(\pi, \mathbf{0}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbb{1}_{t \geq \tau_\pi^{(i)}} \frac{\mathbb{1}_{A_{1:t-1}^{(i)}=0}}{\prod_{t'=1}^{t-1} \hat{e}_{t',0}^{-q(i)}(S_{1:t'}^{(i)})} \hat{\Psi}_{t,W_\pi}(S_{1:t}^{(i)}) \tag{14}$$

for each policy $\pi \in \Pi$, where the relevant DR score is

$$\begin{aligned} \hat{\Psi}_{t,k}(S_{1:t}^{(i)}) = {} & \hat{\mu}_{\text{now},k}^{-q(i)}(S_{1:t}^{(i)}, t) - \hat{\mu}_{\text{next},k}^{-q(i)}(S_{1:t}^{(i)}, t) \\ & + \mathbb{1}_{A_t^{(i)}=k} \frac{Y^{(i)} - \hat{\mu}_{\text{now},k}^{-q(i)}(S_{1:t}^{(i)}, t)}{\hat{e}_{t,k}^{-q(i)}(S_{1:t}^{(i)})} \\ & - \mathbb{1}_{A_t^{(i)}=0} \mathbb{1}_{A_{t+1}^{(i)}=k} \frac{Y^{(i)} - \hat{\mu}_{\text{next},k}^{-q(i)}(S_{1:t}^{(i)}, t)}{\hat{e}_{t,0}^{-q(i)}(S_{1:t}^{(i)}) \hat{e}_{t+1,k}^{-q(i)}(S_{1:t+1}^{(i)})}. \end{aligned} \tag{15}$$

3 Learn the optimal policy by setting $\hat{\pi} = \text{argmax}_{\pi \in \Pi} \hat{\Delta}(\pi, \mathbf{0})$.

[4]Nie et al. Learning when-to-treat policies. JASA. 2021.

Algorithm 2: Advantage doubly robust estimator with terminal state

1 Estimate the outcome models $\mu_{\text{now},k}(\cdot)$, $U(\cdot, \Phi)$, terminating propensities $\rho(\cdot)$ as well as treatment propensities $e_{t,a}(s_{1:t})$ with cross-fitting using any supervised learning method tuned for prediction accuracy.

2 Given these nuisance component estimates, we construct value estimates

$$\hat{\Delta}(\pi, \mathbf{0}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathbb{1}_{S_t \neq \Phi} \mathbb{1}_{t \geq \tau_\pi^{(i)}} \frac{\mathbb{1}_{A_{1:t-1}^{(i)}=0}}{\prod_{t'=1}^{t-1} \hat{e}_{t',0}^{-q(i)}(S_{1:t'}^{(i)})} \hat{\Psi}_{t,W_\pi}^{\Phi}(S_{1:t}^{(i)}) \tag{40}$$

3 for each policy $\pi \in \Pi$, where the relevant DR score is

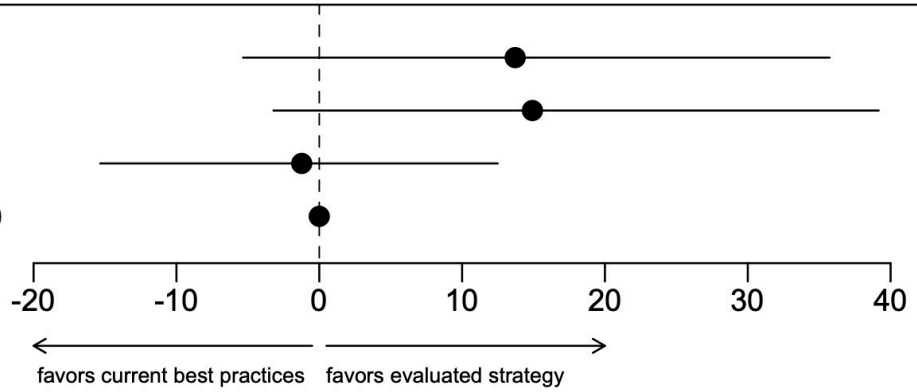
$$\begin{aligned} \hat{\Psi}_{t,k}^{\Phi}(S_{1:t}^{(i)}) = \hat{\mu}_{\text{now},k}^{-q(i)}(S_{1:t}^{(i)}, t) - \hat{\mu}_{\text{next},k}^{\Phi}(S_{1:t}^{(i)}, t)^{-q(i)} \\ + \mathbb{1}_{A_t^{(i)}=k} \frac{Y^{(i)} - \hat{\mu}_{\text{now},k}^{-q(i)}(S_{1:t}^{(i)}, t)}{\hat{e}_{t,k}^{-q(i)}(S_{1:t}^{(i)})} \\ - \mathbb{1}_{A_t^{(i)}=0} \mathbb{1}_{A_{t+1}^{(i)}=k} \frac{Y^{(i)} - \hat{U}^{-q(i)}(S_{1:t}^{(i)}, \Phi)}{\hat{e}_{t,0}^{-q(i)}(S_{1:t}^{(i)}) \hat{e}_{t+1,k}^{-q(i)}(S_{1:t+1}^{(i)})}, \end{aligned} \tag{41}$$

and $\hat{\mu}_{\text{next},k}^{\Phi}(S_{1:t}^{(i)}, t)^{-q(i)} =$
$(1 - \hat{\rho}^{-q(i)}(S_{1:t}^{(i)}))\hat{U}^{-q(i)}(S_{1:t}^{(i)}, \Phi) + \hat{\rho}^{-q(i)}(S_{1:t}^{(i)}) H_t(S_{1:t}^{(i)})$

4 Learn the optimal policy by setting
$\hat{\pi} = \text{argmax}_{\pi \in \Pi} \hat{\Delta}(\pi, \mathbf{0})$.

[4]Nie et al. Learning when-to-treat policies. JASA. 2021.

Evaluated strategy	Mean Difference in HFD60 (95% CI)	Patients initating RRT within three days	Hospital mortality
Crude strategy	13.7 (-5.3 to 35.7)	53%	41%
Stringent strategy	14.9 (-3.2 to 39.2)	14%	38%
Treat all within 24 hours	-1.2 (-15.3 to 12.5)	100%	43%
Standard-delayed (reference)	0 –	38%	45%



	Patient one			Patient two			Patient three		
	First decision timepoint	Second decision timepoint	Third decision timepoint	First decision timepoint	Second decision timepoint	Third decision timepoint	First decision timepoint	Second decision timepoint	Third decision timepoint
Stationary characteristics									
Age (years)	58	58	58	60	60	60	65	65	65
Time-evolving characteristics									
SOFA score	10	10	11	16	15	13	12	12	—
Serum creatinine (mg/dL)	3.6	3.6	3.7	2.1	2.9	3.9	2.2	3.2	—
Blood urea nitrogen (mg/dL)	40	42	47	30	39	53	73	90	—
Serum potassium (mmol/L)	4.8	5.3	4.8	4.2	4.2	4.0	4.6	5.3	—
Arterial blood pH (mmol/L)	7.22	7.25	7.29	7.16	7.20	7.41	7.31	7.31	—
Urine output (ml/kg/min)	0.43	0.37	0.45	0.15	0.10	0.08	0.03	0.01	—
Learned strategy									
Blip (95% CI)	-4.2 (-8.0 to -0.4)	-6.7 (-12.7 to -0.7)	-1.0 (-7.2 to 5.1)	-7.8 (-12.4 to -3.1)	-0.8 (-7.0 to 5.4)	7.2 (0.2 to 14.3)	-5.0 (-9.5 to -0.6)	6.7 (-1.1 to 14.4)	—
Crude strategy's recommendation	Do not initiate	Do not initiate	Do not initiate	Do not initiate	Do not initiate	Initiate*	Do not initiate	Initiate†	Continue

*The stringent strategy would also recommend initiating RRT since the confidence interval shows evidence that patient two will benefit from RRT initiation (i.e., the confidence interval's lower bound is positive).
†Contrary to the crude strategy, the stringent strategy would not recommend initiating RRT since the confidence interval does not show evidence that patient three will benefit from RRT initiation

	MIMIC-III (n=3 748)	AKIKI trials (n=1 068)
Baseline characteristics*		
Age (year)	69 [57–79]	67 [58–75]
Female gender	1 695 (45.2)	344 (32.2)
Weight (kg)	89 [73–107]	81 [69–95]
Non-corticosteroid immunosuppressive drug	62 (1.7)	53 (5.0)
SOFA score (0 to 24)	9 [6–12]	11 [9–13]
Serum creatinine (mg/dL)	1.40 [0.90–2.40]	3.39 [2.57–4.33]
Blood urea nitrogen (mg/dL)	29 [19–47]	56 [39–78]
Serum potassium (mmol/L)	4.2 [3.9–4.7]	4.4 [3.9–5.0]
Arterial blood pH	7.38 [7.33–7.42]	7.31 [7.24–7.37]
Urine output (ml/kg/h)	0.28 [0.22–0.29]	0.12 [0.04–0.34]
Characteristics at H24†		
Blood urea nitrogen (mg/dL)	34 [21–53]	64 [48–90]
Serum potassium (mmol/L)	4.1 [3.8–4.5]	4.4 [3.9–5.0]
Arterial blood pH	7.38 [7.33–7.42]	7.31 [7.25–7.38]
Urine output (ml/kg/h)	0.38 [0.24–0.64]	0.28 [0.08–0.70]
Characteristics at H48‡		
Blood urea nitrogen (mg/dL)	35 [21–56]	67 [48–92]
Serum potassium (mmol/L)	4.1 [3.8–4.4]	4.3 [3.8–4.9]
Arterial blood pH	7.39 [7.34–7.43]	7.34 [7.27–7.40]
Urine output (ml/kg/h)	0.58 [0.31–0.98]	0.41 [0.09–0.88]

Data are n (%) or median [IQR]. IQR=Interquartile range. SOFA score=Sequential Organ Failure Assessment score. To convert the values for creatinine to micrograms per liter, multiply by 88.4. To convert values for blood urea nitrogen to millimoles per litter, multiply by 0.357.
*Characteristics measured just before the first decision timepoint. †Characteristics measured just before the second decision timepoint.
‡Characteristics measured just before the third decision timepoint.

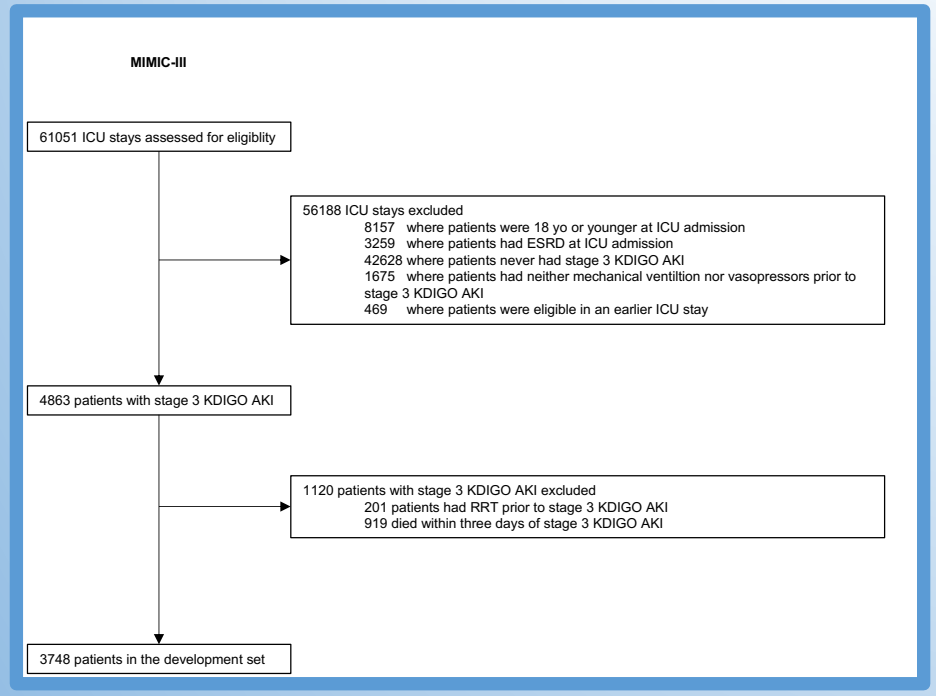
Table 2 Blip parameter estimates from the learned strategies. Estimations based on the multiple imputation analysis of one hundred data sets.

Tailoring covariate	$\hat{\psi}$	(95% CI)
First decision[a]		
Intercept$_1$	-39.589	(-63.885 to -15.294)
Age $_{t=1}$ *(years)*	0.245	(0.035 to 0.454)
Creatinine $_{t=1}$ *(mg/dL)*	1.349	(-0.317 to 3.015)
Potassium $_{t=1}$ *(mmol/L)*	3.409	(-0.547 to 7.364)
Second decision[b]		
Intercept$_2$	-7.747	(-23.343 to 7.849)
SOFA score$_{t=2}$	0.514	(-0.372 to 1.400)
Blood urea nitrogen$_{t=2}$ *(mg/dL)*	0.095	(-0.033 to 0.223)
|pH $_{t=1}$ - pH $_{t=2}$|	-63.874	(-118.998 to -8.750)
Urine output $_{t=1}$ + Urine output $_{t=2}$ *(ml/kg/h)*	-7.734	(-15.303 to -0.165)
Third decision[c]		
Intercept$_3$	5.397	(-14.443 to 25.237)
Urine output $_{t=3}$ *(ml/kg/h)*	-19.316	(-34.365 to -4.268)
Blood urea nitrogen $_{t=3}$/Blood urea nitrogen $_{t=1}$	1.922	(-10.974 to 14.818)

Denoting H_t a patient's vector of covariates at decision timepoint t; $\widehat{M}_t$ the estimated variance-covariance matrix from decision timepoint t; $\hat{\psi}_t$ the blip parameter estimates from decision timepoint t, 95% confidence intervals for the individual blips can be calculated as

$$\hat{\psi}_t{}^T H_t \pm 1.96 \times \sqrt{H_t{}^T \widehat{M}_t H_t}.$$

Decision point	Intercept	First variable	Second variable	Third variable	Fourth variable
First decision	Intercept_1	$\text{Age}_{t=1}$	$\text{Creatinine}_{t=1}$	$\text{Potassium}_{t=1}$	—
Intercept_1	153.66	-0.843	-0.920	-20.615	
$\text{Age}_{t=1}$	-0.843	0.011	-0.005	0.039	
$\text{Creatinine}_{t=1}$	-0.920	-0.005	0.722	-0.263	
$\text{Potassium}_{t=1}$	-20.615	0.039	-0.263	4.073	
Second decision	Intercept_2	$\text{SOFA score}_{t=2}$	$\text{Blood urea nitrogen}_{t=2}$	$\lvert \text{pH}_{t=1} - \text{pH}_{t=2} \rvert$	$\text{Urine output}_{t=1} + \text{Urine output}_{t=2}$
Intercept_2	63.319	-2.711	-0.270	-74.776	-9.185
$\text{SOFA score}_{t=2}$	-2.711	0.204	0.001	1.934	0.174
$\text{Blood urea nitrogen}_{t=2}$	-0.270	0.001	0.004	0.077	-0.022
$\lvert \text{pH}_{t=1} - \text{pH}_{t=2} \rvert$	-74.776	1.934	0.077	791.019	1.372
$\text{Urine output}_{t=1} + \text{Urine output}_{t=2}$	-9.185	0.174	-0.022	1.372	14.914
Third decision	Intercept_3	$\text{Urine output}_{t=3}$	$\text{Blood urea nitrogen}_{t=3}$ / $\text{Blood urea nitrogen}_{t=1}$		—
Intercept_3	102.467	-23.944	-63.053		
$\text{Urine output}_{t=3}$	-23.944	58.95	5.045		
$\text{Blood urea nitrogen}_{t=3}$ / $\text{Blood urea nitrogen}_{t=1}$	-63.053	5.045	43.292		


MIMIC-III
61051 ICU stays assessed for eligiblity
56188 ICU stays excluded
8157 where patients were 18 yo or younger at ICU admission
3259 where patients had ESRD at ICU admission
42628 where patients never had stage 3 KDIGO AKI
1675 where patients had neither mechanical ventiltion nor vasopressors prior to stage 3 KDIGO AKI
469 where patients were eligible in an earlier ICU stay
4863 patients with stage 3 KDIGO AKI
1120 patients with stage 3 KDIGO AKI excluded
201 patients had RRT prior to stage 3 KDIGO AKI
919 died within three days of stage 3 KDIGO AKI
3748 patients in the development set

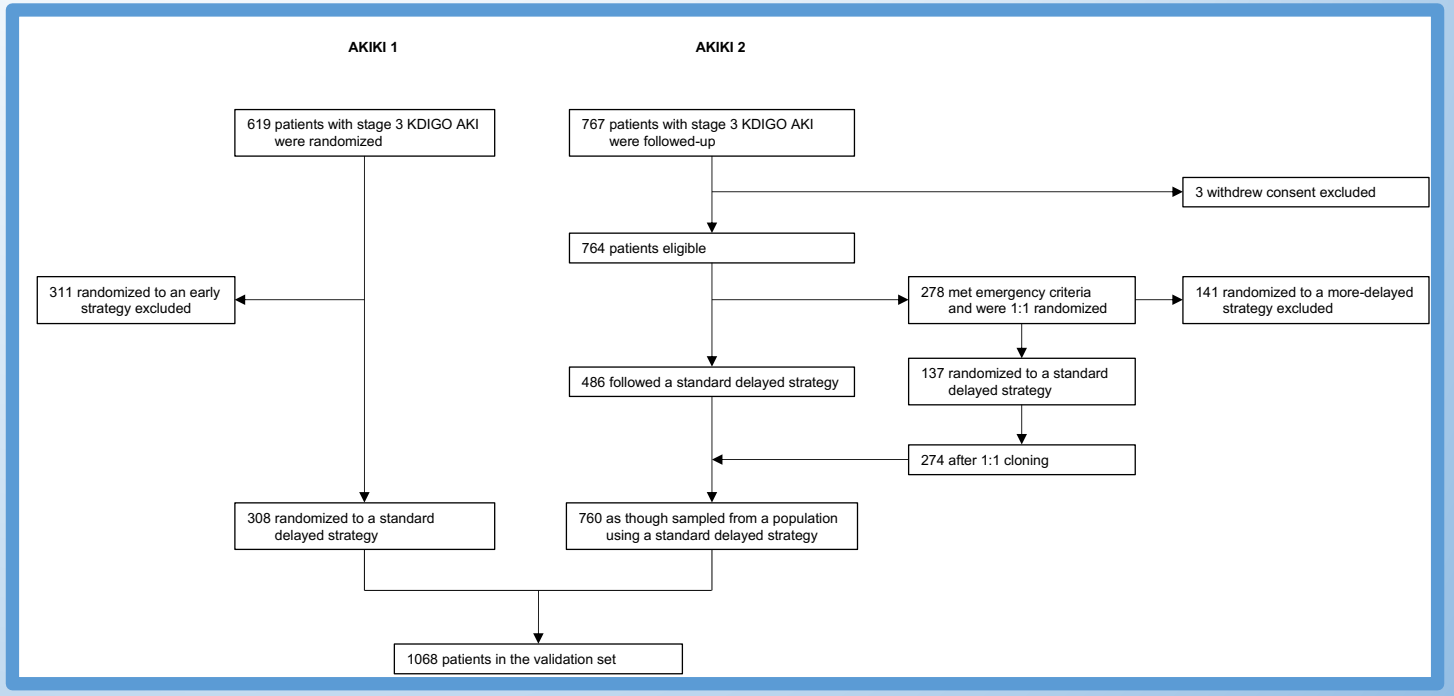
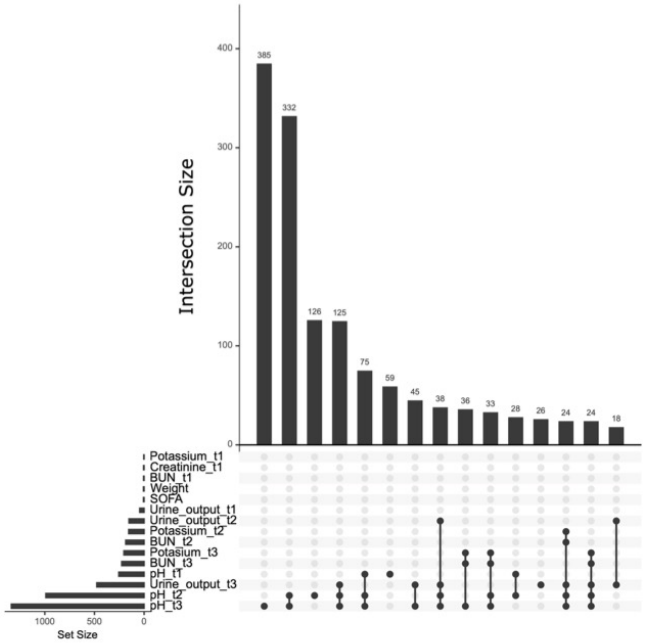

AKIKI 1
AKIKI 2
619 patients with stage 3 KDIGO AKI were randomized
767 patients with stage 3 KDIGO AKI were followed-up
3 withdrew consent excluded
764 patients eligible
311 randomized to an early strategy excluded
278 met emergency criteria and were 1:1 randomized
141 randomized to a more-delayed strategy excluded
486 followed a standard delayed strategy
137 randomized to a standard delayed strategy
274 after 1:1 cloning
308 randomized to a standard delayed strategy
760 as though sampled from a population using a standard delayed strategy
1068 patients in the validation set

Figure S3. Missing data patterns in the development set (A) and validation set (B).

A



B