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

François GROLLEAU,[1] François PETIT*,[2] Stéphane GAUDRY*,[3] Élise DIARD,[4] Jean-Pierre QUENOT,[5] Didier DREYFUSS,[6] Viet-Thi TRAN,[7] Raphaël PORCHER.[8]

* These authors contributed equally as second authors.

Manuscript words count: 3497

Source of Funding and Conflicts of Interest:

François Petit was supported by a "Chaire d'excellence" (excellence fellowship) from the IdEx Université Paris Cité, ANR-18-IDEX-0001. Raphaël Porcher acknowledges the support of the French Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). The authors have disclosed that they do not have any conflicts of interest.

Corresponding author:

François Grolleau

Hôtel-Dieu Hospital, 1 place du Parvis Notre-Dame 75004 Paris, France.

Email: francois.grolleau@aphp.fr

Tel: +33 1 42 34 89 87

[1] M.D., M.P.H., Ph.D. Candidate, Assistant Professor, Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Center for Research in Epidemiology and StatisticS (CRESS), and Centre d'Epidémiologie Clinique, AP-HP, Hôpital Hôtel Dieu, F-75004 Paris, France.

[2] Ph.D., Junior Professor, Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Center for Research in Epidemiology and StatisticS (CRESS), F-75004 Paris, France.

[3] M.D., Ph.D., Professor, AP-HP, Hôpital Avicenne, Service de Réanimation Médico-Chirurgicale, Université Paris 13, Bobigny, Health Care Simulation Center, UFR SMBH, Sorbonne Paris Cité and INSERM UMR S1155 "Common and Rare Kidney Diseases: from Molecular Events to Precision Medicine," Paris, France.

[4] M.S., Information technology developer, Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Center for Research in Epidemiology and StatisticS (CRESS), and Centre d'Epidémiologie Clinique, AP-HP, Hôpital Hôtel Dieu, F-75004 Paris, France.

[5] M.D., Ph.D., Professor, Department of Intensive Care, François Mitterrand University Hospital, Lipness Team, INSERM Research Center, LNC-UMR1231 and LabEx LipSTIC, and INSERM CIC 1432, Clinical Epidemiology, University of Burgundy, Dijon, France.

[6] M.D., Professor Emeritus, Université Paris Cité, Service de Médecine Intensive-Réanimation, Hôpital Louis Mourier, AP-HP and INSERM UMR S1155 "Common and Rare Kidney Diseases: from Molecular Events to Precision Medicine," Sorbonne Université, Paris, France.

[7] M.D., Ph.D., Associate Professor, Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Center for Research in Epidemiology and StatisticS (CRESS), and Centre d'Epidémiologie Clinique, AP-HP, Hôpital Hôtel Dieu, F-75004 Paris, France.

[8] Ph.D., Professor, Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Center for Research in Epidemiology and StatisticS (CRESS), and Centre d'Epidémiologie Clinique, AP-HP, Hôpital Hôtel Dieu, F-75004 Paris, France.

Abstract

Background: Trials sequentially randomizing patients each day have never been conducted for renal replacement therapy (RRT) initiation. We used clinical data from routine care and trials to learn and validate optimal dynamic strategies for RRT initiation in the intensive care unit (ICU).

Methods: We included participants from the MIMIC-III database for development, and AKIKI and AKIKI2 (two randomized controlled trials on RRT timing) for validation. Participants were eligible if they were adult ICU patients with severe acute kidney injury, receiving invasive mechanical ventilation, catecholamine infusion, or both. We used doubly-robust estimators to learn when to start RRT after the occurrence of severe acute kidney injury given a patient's evolving characteristics—for three days in a row. The 'crude strategy' aimed to maximize hospital-free days at day 60 (HFD60). The 'stringent strategy' recommended initiating RRT only when there was evidence at the 0.05 threshold that a patient would benefit from initiation. For external validation, we evaluated the causal effects of implementing our learned strategies *versus* following current best practices on HFD60.

Results: We included 3 748 patients in the development set (median age 69y [IQR 57–79], median SOFA score 9 [IQR 6–12], 1 695 [45.2%] female), and 1 068 in the validation set (median age 67y [IQR 58–75], median SOFA score 11 [IQR 9–13], 344 [32.2%] female). Through external validation, we found that compared to current best practices, the crude and stringent strategies improved average HFD60 by 13.7 [95% CI -5.3–35.7], and 14.9 [95% CI -3.2–39.2] days respectively. Contrasted to current best practices where 38% of patients initiated RRT within three days, with the stringent strategy, we estimated that only 14% of patients would.

Conclusion: We developed a practical and interpretable dynamic decision support system for RRT initiation in the ICU. Its implementation could improve the average number of days that ICU patients spend alive and outside the hospital.

Key words acute kidney injury, renal replacement therapy, personalized medicine, causal inference, reinforcement learning.

Introduction

In intensive care units (ICU), acute kidney injury (AKI) affects about one in two patients, and its onset is associated with high mortality and long-term sequelae (1). Renal replacement therapy (RRT) is an invasive but potentially life-saving treatment for AKI (2). Because AKI is a heterogeneous and rapidly evolving syndrome (3), controversies on the timing and selection of patients for initiating RRT have long prevailed (4). In the last decade, multicenter randomized trials, compared early versus delayed RRT initiation strategies, but the analyses (5–7) and meta-analyzes (8, 9) of these trials failed to show significant differences in patient-important outcomes at the population level. As such negative trial findings are widespread in critical care, identifying individualized treatment effects has been judged a research priority (10).

Physicians' attempts to deliver timely interventions tailored to patients' characteristics have a long history (11). While in some diseases, biological insight proved decisive in moving precision medicine forward (12), AKI—due to its heterogeneous syndromic nature—is less amenable to this approach. Recently, authors proposed algorithms for RRT initiation in the ICU (13, 14), but the need for validated data-driven decision support tools remains (15). Previously, we developed a decision support tool based on clinical trial data and considered the static case where the decision to initiate RRT is only pondered at AKI onset (16). Yet, for such decision tools to be actionable and consistent with practice, they must go beyond the static case and account for the fundamentally dynamic nature of AKI. In fact, when a decision support tool recommends not initiating RRT for a given patient on a given day, it ought to re-evaluate its recommendation on the next day considering the evolution of the patient's characteristics.

To learn an optimal RRT initiation strategy under this setting, the ideal method would be to conduct a Sequential Multiple Assignment Randomized Trial (SMART) where AKI patients are sequentially randomized each day to either initiate treatment or not (17). Due to cost, time, and practical constraints, SMART trials have never been conducted in the ICU. However, recent developments in statistics and computer science provided robust methods to learn and evaluate optimal treatment initiation strategies from observational data (18–20). To our knowledge, only a single monocenter study has analyzed clinical data in an attempt to develop a dynamic decision support system for RRT initiation (21).

In this paper, we used reinforcement learning methods on data from electronic health records to estimate optimal dynamic strategies for RRT initiation in ICU patients with AKI. Then, in an external validation step, we used data from two large multicenter randomized trials on RRT timing to estimate the benefit of implementing these strategies.

Methods

Sources of data

The development sample included participants from the Multi-Parameter Intelligent Monitoring in Intensive Care III (MIMIC-III) database. MIMIC-III is a project maintained by the Laboratory for Computational Physiology at the Massachusetts Institute of Technology which contains routinely collected data from 61,051 distinct ICU admissions of adult patients admitted between 2001 and 2012 (22). For reproducibility, we used the database's official code repository to extract all relevant variables (23). As out-of-hospital mortality was not available in the latest version of the MIMIC project, we used MIMIC-III version 1.4.

The validation sample included participants from the AKIKI and AKIKI2 trials, two multicenter RCTs conducted in France (5, 24). The AKIKI trial was conducted at 31 ICUs from Sept 2013 through Jan 2016 and recruited 619 patients with stage 3 KDIGO-AKI who required mechanical ventilation, catecholamine infusion, or both. Included patients were 1:1 randomized to either an early RRT initiation strategy or to a standard-delayed initiation strategy. The AKIKI2 trial was embedded in a cohort recruiting at 39 ICUs from May 2018 through Oct 2019. In AKIKI2, eligibility criteria for the cohort were identical to the eligibility criteria from the original AKIKI trial. Of the 767 patients included in the cohort, 278 met one or more randomization criteria (oliguria for more than 72h or blood urea nitrogen concentration greater than 112 mg/dL) and were 1:1 allocated to either a standard-delayed RRT initiation strategy or to a more-delayed strategy.

Population

Eligible patients were adults (18 years of age or older) hospitalized in the ICU with stage 3 KDIGO AKI who were receiving (or had received for this episode) invasive mechanical ventilation, catecholamine infusion, or both. Staging in the KDIGO classification was based on serum creatinine and/or urine output with higher stages indicating greater severity (25). As

the latest clinical guidelines recommend a standard-delayed strategy of RRT initiation (26), we chose this strategy as the reference "best practice" upon which to improve. Precisely, our target population was made of individuals whose physicians implemented a standard-delayed strategy. In both AKIKI trials, the standard-delayed strategy suggested initiating RRT if one of the following criteria occurred: severe hyperkalemia and/or metabolic acidosis, pulmonary oedema resistant to diuretics, oliguria for more than 72 hours, blood urea nitrogen level higher than 112 mg per deciliter. In the current study, we used the same exclusion criteria as in the AKIKI trials i.e., moribund state (patient likely to die within 72h), end-stage kidney disease (i.e., patient with creatinine clearance < 15ml/ml), patients having received RRT before inclusion, and patients already included at a previous date.

Setup and timepoints for learning dynamic RRT initiation strategies

From a clinical standpoint, the decision to start RRT is considered difficult in the first days following severe AKI. After three days, this decision often becomes straightforward, as most patients have either recovered or deteriorated. We focused on developing a when-to-treat strategy for RRT initiation in the first 72 hours following the onset of severe AKI (i.e., stage 3 KDIGO-AKI). Specifically, we learned a strategy that—for three days in a row after severe AKI onset—assessed the need to start RRT given a patient's evolving characteristics. Our strategy was non-stationary i.e., the decision rules for RRT initiation could differ depending on the day. We considered three decision timepoints at 0, 24, and 48 hours after severe AKI onset (**Figure 1**). The strategy was developed so that, at each timepoint, it used clinical and biological information gathered prior to this timepoint as inputs and outputted a recommendation to either initiate RRT within 24 hours or not. We considered that once RRT had been recommended (or initiated in contradiction with the strategy's recommendation), the strategy would persist in recommending RRT for all subsequent decision timepoints. This so-called regularity in the strategy's behavior indicates that we did not consider when to stop RRT in the three days following severe AKI onset.

Primary outcome

The primary outcome was hospital-free days at day 60 (HFD60). This outcome was chosen because i) it was a good compromise between patient-centeredness and pragmatism (27); and ii) it reduced the risk that the learned strategy had unexpected side effects—a well-known issue in reinforcement learning (28). For instance, using short-term mortality as a primary outcome,

the model may learn a strategy that maximizes survival at the cost of keeping patients alive in the ICU as long as possible.

Learning an optimal strategy

To learn an optimal strategy, we used a doubly robust estimator with weighted least squares (dWOLS) (29). This method relies on estimating blip functions for each decision timepoint. Given the evolving characteristics of an individual up to timepoint t, a blip function predicts the effect of initiating RRT at t versus not initiating it at t but taking optimal treatment decisions from timepoint $t+1$ onwards. We derived two strategies from the estimated blip functions. We termed "crude" the strategy that recommended RRT initiation to the patients with positive values of blips, and "stringent" the strategy that recommended initiating RRT only when there was evidence at the 0.05 significance level that a patient would benefit from RRT initiation (i.e., positive lower bound for the blip's 95% confidence interval). As stated before, once RRT initiation was recommended, both strategies persisted in recommending RRT regardless of the blips at subsequent timepoints. Patients who died within three days of AKI were excluded from the development sample, considering no relevant information would be learned from these patients' data. Indeed, it seemed unlikely that patients who died within three days of severe AKI could have been discharged from the hospital under a different RRT initiation strategy: we expected their outcome to be the same under all strategies (HFD60 is zero for all patients who die in the hospital). However, these patients were not excluded from the validation sample, to avoid time-dependent selection bias. More details on dWOLS estimation and inference are given in the appendix (pp 3-4).

External validation

To match our target "best practice" population, we included all patients from AKIKI and AKIKI2 who had received a standard-delayed strategy. As the AKIKI2 patients randomized to a more-delayed strategy were compatible with a standard-delayed strategy until they met a randomization criterion, we excluded these patients but duplicated the patients randomized to the standard-strategy arm according to the cloning and censoring principle used for emulating target trials from observational data (30). To estimate hospital mortality and the proportion of patients who would initiate RRT within three days under a strategy, we used importance sampling for policy evaluation in reinforcement learning (31).

To evaluate the effect of new strategies on HFD60, we considered current best practices (i.e., the standard-delayed strategy from the AKIKI trials) as a common control and compared

it to the following three strategies: i) the crude strategy, ii) the stringent strategy, and iii) a strategy that recommends initiating RRT in all patients within 24 hours after severe AKI onset. We estimated the causal effect of implementing each of these strategies compared to current best practices using the cross-fitted advantage doubly robust estimator for strategy evaluation with terminal states (19). This estimator allows estimating the mean difference in the outcome that would have been observed under any given strategy and the outcome observed under a reference strategy. We provide more details on importance sampling for policy evaluation and the advantage doubly robust estimator in the appendix (pp 4-5).

Ethical approval and research transparency

The MIMIC-III analysis received approval from the Institutional Review Boards of the Massachusetts Institute of Technology and Beth Israel Deaconess Medical Center (BIDMC). The AKIKI and AKIKI2 trials received approval from competent French legal authority (Comité de Protection des Personnes d'Ile de France VI, ID RCB 2013-A00765-40, NCT01932190 for AKIKI and ID RCB 2017-A02382-51, NCT03396757 for AKIKI 2) and participants provided written informed consent to take part in the study. The funding sources were not involved in the study design; collection, analysis, and interpretation of data; writing of the manuscript; or the process of submission for publication. Two authors (FG, RP) had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the analysis. Analyses were conducted using R version 4.2.1 for strategy learning as well as plotting, and Python 3.8.8 for strategy evaluation. The code used in this study is available at https://github.com/fcgrolleau/dynamic-rrt.

Results

Learning optimal dynamic strategies for RRT initiation

1. Patients

From 2001 through 2012, a total of 3 748 ICU patients with AKI recruited at a tertiary teaching hospital (BIDMC — Harvard Medical School) met eligibility criteria and were included in the development set (**Figure S1, Panel A**). Almost half of individuals were females (n=1 695; 45.2%). At enrollment, patients had a mean SOFA score of 9 (interquartile range [IQR], 6–12). All patients had severe AKI (i.e., stage 3 KDIGO-AKI) which diagnosis was most often based on urine output (n=3 328; 88.8%). At enrollment, the median serum creatinine and urine output were 1.40 mg/dL (IQR, 0.90–2.40) and 0.28 ml/kg/h (IQR, 0.22–0.29) respectively. During the follow-up, 400 (10.7%) patients initiated RRT within three days of severe AKI, and 892 (23.8%) died during hospitalization. The mean and median HDF60 were 33.6 and 42.9 days, (IQR, 0.9–51.7) respectively. Additional baseline and evolving characteristics for these patients are given in **Table 1**.

2. Learned strategies

For patients with severe AKI who have never initiated RRT at a given decision timepoint, decision rules whether to initiate RRT in the next 24 hours were derived from the models we estimated at each timepoint (**Figure 1**). Estimated parameters of the so-called blip functions are given with didactical instructions for calculations in **Table 2 (**their covariances are given in **Table S1**). In **Figure 2A** we display the recommendations from two learned strategies (i.e., our crude and stringent strategies) along with the uncertainty in the recommendation for each patient in the development set. We present in **Table 3**, three illustrative examples where the learned strategies were applied for individualizing the decision to initiate RRT within 72 hours of severe AKI. The apparent effect (i.e., in the development set) of implementing our crude strategy versus implementing the MIMIC-III RRT initiation strategy was a 6.6 days improvement in mean HFD60.

External validation

1. Patients

From 2013 through 2019, a total of 931 unique ICU patients with AKI from the AKIKI and AKIKI2 trials met our predefined eligibility criteria and were included in the validation set. After cloning and censoring, these corresponded to a sample of 1 068 individuals from a population who have received current best practices (i.e., a standard-delayed strategy, *see* **Figure S1, Panel B**). About a third of individuals (n=344; 32.2%) from the validation set were females. For most patients, severe AKI was associated with septic shock and the mean SOFA score was 11 (IQR, 9–13). A drop in urine output triggered stage 3 KDIGO-AKI diagnosis in 401 patients (37.5%). At enrolment, the median serum creatinine and urine output were 3.39 mg/dL (IQR, 2.57–4.33) and 0.12 ml/kg/h (IQR, 0.04–0.34) respectively. During follow-up, 482 (45.1%) died during hospitalization, while 405 (38%) and 99 (20.5%) respectively initiated RRT or died within three days of severe AKI. The mean HFD60 was 14.2 days (median 0, IQR, 0–30.3).

2. External validation of the learned strategies

In the external validation population, we estimated that, under our crude strategy, 41% of patients would die during hospitalization and 53% would initiate RRT within three days. Under our stringent strategy, we estimated that 38% of patients would die during hospitalization and 14% of patients would initiate RRT within three days. Recommendations from the learned strategies along with the uncertainty in individual-patient recommendations are given for all patients in the validation set in **Figure 2B**. The discrepancies between current best practices and the recommendations from the learned strategies are shown in **Figure 3**. We found that compared to current best practices (i.e., the standard-delayed strategy from the AKIKI trials), our crude and stringent strategies yielded a 13.7 days and 14.9 days improvement in mean HFD60 respectively (**Figure 4**).

Discussion

Summary of findings

In this study, we used electronic health record data to learn dynamic RRT initiation strategies for ICU patients with severe AKI. Then, using data from two large RCTs of RRT timing we conducted external validation: compared to current best practices (i.e., a standard-delayed strategy), we found that the crude strategy may improve HFD60 by 13.7 days on average. Note that even though the crude strategy may recommend RRT initiation sooner than the standard-delayed strategy, it is not an early strategy. Consistent with previous trials (5–7), we showed that a strategy that recommends RRT initiation in all patients within 24 hours of severe AKI may yield outcomes similar to or worse than that of a standard-delayed strategy. In contrast to early strategies, the crude strategy identified that only 53% of patients required RRT initiation in the three days following severe AKI. Of note, in the STARRT-AKI and AKIKI arms corresponding to current best practices (i.e., the arms termed standard and delayed respectively), rates of RRT initiations a week after severe AKI were 59% and 55%. We believe that the benefit of the crude strategy stems from its ability to identify earlier the patients who will ultimately require RRT. That said, we found that the stringent strategy may also improve patients' HFD60 all the while reducing RRT prescriptions in the three days following severe AKI. This suggests that the individual-patient confidence intervals given by the crude strategy provide important information for deciding the initiation of RRT. Entailing less frequent usage of RRT, the stringent strategy could have the benefit of not only improving patient-important outcomes but also saving health resources.

Our methodology aimed at developing interpretable linear decision rules together with confidence intervals to guide clinicians at the bedside. For greater transparency and interpretability, we released a user-friendly online implementation of our learned strategies at http://dynamic-rrt.eu. Using the time-varying characteristics of a patient as input, clinicians can with this web application obtain individual-patient recommendations from the crude strategy along its 95% confidence intervals. With respect to interpretability, we noticed that on the first day, the crude strategy recommended RRT initiation more often in older patients with higher values of serum creatinine and serum potassium. On the second day, it seemed inclined to recommend RRT initiation in patients with stable arterial pH having a critical combination of low urine output and high blood urea nitrogen levels. Only on the third day did the learned strategies appear more aggressive recommending RRT initiation in most patients who had not

recovered kidney function (i.e., patients with persisting low urine output and high blood urea nitrogen levels).

In this work, we chose to use HFD60 rather than mortality as the primary outcome. Mortality at a given timepoint conveys limited statistical information, as it contains only two possible values. In recent years, there has been an increased focus on patient-cantered non-mortality outcomes such as event-free day endpoints in ICU research (27, 32). In a dynamic reinforcement learning setting, there is however one more reason not to use survival as the outcome to optimize. Using survival as a distal reward signal may push the system to find a strategy that maximizes survival at the cost of unnecessary invasive procedures. Practically, the model could use its many degrees of freedom to learn a strategy that increases 60-day survival but decreases hospital and ICU discharge. On the contrary, optimizing over HFD60 is unlikely to yield longer hospital or ICU stays (33).

Strength and limitations

To our knowledge, this study is the first to provide a validated dynamic decision support system for RRT initiation in the ICU. We believe the implications of our work are not only clinical but also methodological as the approach we used can be adapted for the timely initiation of a wide variety of treatments in medicine. However, our study has serval limitations. First, we considered only regular strategies, i.e., we did not allow for strategies to recommend stopping RRT before the third day if it had been initiated earlier. Disregarding the opportunities to stop treatment had a strong statistical advantage as it decreased the opportunities for a mismatch between prescribed and recommended treatments, thereby reducing variance in strategy learning and strategy evaluation. From a clinical standpoint, finding an optimal stopping strategy would rather be a distinct question that is more relevant after the third day. Second, we acknowledge that the effect size from implementing our learned strategies, though clinically relevant, was not statistically significant at the conventional 0.05 threshold. In reinforcement learning, learned strategies have long been tested on their training data, and inference for strategy evaluation is still rarely provided as reaching statistical significance often requires huge sample sizes (34). In this study, we performed external validation and estimated confidence intervals of the strategies' benefits. This transparent approach indicates that developing more robust strategies may require training and testing on larger databases, perhaps coupling multiple electronic health records. Third, we concede that given infinitely large sample sizes, methods that leverage computation rather than expert knowledge (e.g., methods such as deep Q networks) may ultimately be more effective. Nevertheless, we believe that as

even large electronic health records yield small effective sample sizes, encoding expert knowledge in the feature engineering process remains essential. Compared to black-box algorithms, we trust this human-centric approach is more likely to convince clinicians as it offers a window for interpretability.

Implication for future research

As is true of traditional drugs, new individualized strategies will require proper testing in clinical settings before they can be deployed (35). This could be done for instance in a cluster randomized controlled trial comparing physicians alone to physicians assisted by the clinical decision support system. Alternatively, new trial designs could help to improve the learned strategy while it is being prospectively evaluated (36). Finally, if kidney damage markers (e.g., C-C motif chemokine ligand 14) demonstrate their clinical utility (37), new strategies leveraging this information may be developed. In the long run, these developments may help bridge the gap between biological knowledge and actionable data-driven approaches. We believe that fostering collaborations of clinical experts, methodologists, and mathematicians all genuinely interested in AKI and reinforcement learning is key. This, we hope, will continue to move personalized medicine forward for the benefit of intensive care patients.

In conclusion, we developed a dynamic RRT initiation strategy and confirmed via external validation that its implementation could increase the average number of days that ICU patients spend alive and outside the hospital. This interpretable strategy relies on routinely collected data and provides confidence intervals to guide decision-making at the bedside. It will require prospective testing and refinements before it can be broadly deployed in practice.

ABREVIATIONS

AKI: Acute Kidney Injury

BIDMC: Beth Israel Deaconess Medical Center

CI: Confidence Interval

HFD60: Hospital-Free Days at day 60

ICU: Intensive Care Unit

IQR: Interquartile Range

KDIGO: Kidney Disease: Improving Global Outcomes

RCT: Randomized Controlled Trial

RRT: Renal Replacement Therapy

SMART: Sequential Multiple Assignment Randomized Trial

SOFA: Sequential Organ Failure Assessment

References

1. Hoste EAJ, Bagshaw SM, Bellomo R, Cely CM, Colman R, Cruz DN, *et al.* Epidemiology of acute kidney injury in critically ill patients: the multinational AKI-EPI study. *Intensive Care Med* 2015;41:1411–1423.
2. Gaudry S, Palevsky PM, Dreyfuss D. Extracorporeal Kidney-Replacement Therapy for Acute Kidney Injury. *N Engl J Med* 2022;386:964–975.
3. Ronco C, Bellomo R, Kellum JA. Acute kidney injury. *Lancet* 2019;394:1949–1964.
4. Ostermann M, Bellomo R, Burdmann EA, Doi K, Endre ZH, Goldstein SL, *et al.* Controversies in acute kidney injury: conclusions from a Kidney Disease: Improving Global Outcomes (KDIGO) Conference. *Kidney Int* 2020;98:294–309.
5. Gaudry S, Hajage D, Schortgen F, Martin-Lefevre L, Pons B, Boulet E, *et al.* Initiation Strategies for Renal-Replacement Therapy in the Intensive Care Unit. *N Engl J Med* 2016;375:122–133.
6. Barbar SD, Clere-Jehl R, Bourredjem A, Hernu R, Montini F, Bruyère R, *et al.* Timing of Renal-Replacement Therapy in Patients with Acute Kidney Injury and Sepsis. *N Engl J Med* 2018;379:1431–1442.
7. STARRT-AKI Investigators, Canadian Critical Care Trials Group, Australian and New Zealand Intensive Care Society Clinical Trials Group, United Kingdom Critical Care Research Group, Canadian Nephrology Trials Network, Irish Critical Care Trials Group, *et al.* Timing of Initiation of Renal-Replacement Therapy in Acute Kidney Injury. *N Engl J Med* 2020;383:240–251.
8. Fayad AII, Buamscha DG, Ciapponi A. Timing of renal replacement therapy initiation for acute kidney injury. *Cochrane Database Syst Rev* 2018;12:CD010612.
9. Gaudry S, Hajage D, Benichou N, Chaïbi K, Barbar S, Zarbock A, *et al.* Delayed versus early initiation of renal replacement therapy for severe acute kidney injury: a systematic review and individual patient data meta-analysis of randomised clinical trials. *Lancet* 2020;395:1506–1515.
10. Semler MW, Bernard GR, Aaron SD, Angus DC, Biros MH, Brower RG, *et al.* Identifying Clinical Research Priorities in Adult Pulmonary and Critical Care. NHLBI Working Group Report. *Am J Respir Crit Care Med* 2020;202:511–523.
11. Phillips CJ. Precision Medicine and Its Imprecise History. *Harvard Data Science Review* 2020;2:.

12. Romond EH, Perez EA, Bryant J, Suman VJ, Geyer CE, Davidson NE, *et al.* Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *N Engl J Med* 2005;353:1673–1684.
13. Gaudry S, Quenot J-P, Hertig A, Barbar SD, Hajage D, Ricard J-D, *et al.* Timing of Renal Replacement Therapy for Severe Acute Kidney Injury in Critically Ill Patients. *Am J Respir Crit Care Med* 2019;199:1066–1075.
14. Bagshaw SM, Hoste EA, Wald R. When should we start renal-replacement therapy in critically ill patients with acute kidney injury: do we finally have the answer? *Critical Care* 2021;25:179.
15. Schaub JA, Heung M. Precision Medicine in Acute Kidney Injury: A Promising Future? *Am J Respir Crit Care Med* 2019;199:814–816.
16. Grolleau F, Porcher R, Barbar S, Hajage D, Bourredjem A, Quenot J-P, *et al.* Personalization of renal replacement therapy initiation: a secondary analysis of the AKIKI and IDEAL-ICU trials. *Critical Care* 2022;26:64.
17. Almirall D, Nahum-Shani I, Sherwood NE, Murphy SA. Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research. *Transl Behav Med* 2014;4:260–274.
18. Khezeli K, Siegel S, Shickel B, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. Reinforcement Learning for Clinical Applications. *Clin J Am Soc Nephrol* 2023;18:521–523.
19. Nie X, Brunskill E, Wager S. Learning when-to-treat policies. *Journal of the American Statistical Association* 2021;116:392–409.
20. Tsiatis AA, Davidian M, Holloway ST, Laber EB. *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. CRC Press; 2019.
21. Morzywołek P, Steen J, Vansteelandt S, Decruyenaere J, Sterckx S, Van Biesen W. Timing of dialysis in acute kidney injury using routinely collected data and dynamic treatment regimes. *Crit Care* 2022;26:365.
22. Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:160035.
23. Johnson AEW, Stone DJ, Celi LA, Pollard TJ. The MIMIC code repository: enabling reproducibility in critical care research. *J Am Med Inform Assoc* 2018;25:32–39.
24. Gaudry S, Hajage D, Martin-Lefevre L, Lebbah S, Louis G, Moschietto S, *et al.* Comparison of two delayed strategies for renal replacement therapy initiation for severe acute kidney injury (AKIKI 2): a multicentre, open-label, randomised, controlled trial. *The Lancet* 2021;397:1293–1300.

25. Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group. KDIGO clinical practice guideline for acute kidney injury. *Kidney Int Suppl* 2012;2:1–138.
26. Evans L, Rhodes A, Alhazzani W, Antonelli M, Coopersmith CM, French C, *et al.* Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock 2021. *Crit Care Med* 2021;49:e1063–e1143.
27. Auriemma CL, Taylor SP, Harhay MO, Courtright KR, Halpern SD. Hospital-free days: a pragmatic and patient-centered outcome for trials among critically and seriously ill patients. *Am J Respir Crit Care Med* 2021;204:902–909.
28. Sutton RS, Barto AG. 17.4 Designing reward signals. *Reinforcement learning: An introduction* MIT press; 2018.
29. Wallace MP, Moodie EEM. Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics* 2015;71:636–644.
30. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol* 2016;183:758–764.
36. Precup D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series* 2000; p. 80.
32. Harhay MO, Casey JD, Clement M, Collins SP, Gayat É, Gong MN, *et al.* Contemporary strategies to improve clinical trial design for critical care research: insights from the First Critical Care Clinical Trialists Workshop. *Intensive Care Med* 2020;46:930–942.
33. Hadfield-Menell D, Russell SJ, Abbeel P, Dragan A. Cooperative inverse reinforcement learning. *Advances in neural information processing systems* 2016;29:.
34. Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, *et al.* Guidelines for reinforcement learning in healthcare. *Nat Med* 2019;25:16–18.
35. Komorowski M. Clinical management of sepsis can be improved by artificial intelligence: yes. *Intensive Care Med* 2020;46:375–377.
36. Klasnja P, Hekler EB, Shiffman S, Boruvka A, Almirall D, Tewari A, *et al.* Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychol* 2015;34S:1220–1228.
37. Ostermann M, Zarbock A, Goldstein S, Kashani K, Macedo E, Murugan R, *et al.* Recommendations on Acute Kidney Injury Biomarkers From the Acute Disease Quality Initiative Consensus Conference: A Consensus Statement. *JAMA Netw Open* 2020;3:e2019209.

Acknowledgements

The authors thank Cynthia T. Chen (Westaf) for editing. We thank all patients included in the AKIKI trials as well as their surrogates. We express our gratitude to the medical and nursing teams that participated in these trials.

Authors' contributions

FG, RP, FP, and VTT conceived the study. FG wrote the codes and did the computational analysis with input from RP and FP. SG, JPQ, and DD provided data from the AKIKI trials. ED designed the Sankey diagrams and contributed to the user interface. FG drafted the manuscript with inputs from RP, VTT, FP, SG, DD, and JPQ. All the authors read the paper and suggested edits. RP supervised the project. FG and RP accessed and verified the data. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Competing interests

The authors have disclosed that they do not have any conflicts of interest.

ADDITIONAL INFORMATION

Supplementary information

The online version contains supplementary material available at https://doi.org/XX.

Supplementary Methods: Appendix A Setup notations. **Appendix B** Summary of notations introduced in the appendix. **Appendix C** Doubly robust dynamic treatment regimen via weighted least squares. **Appendix D** Variable selection. **Appendix E** Missing data management. **Appendix E** Importance sampling for policy evaluation. **Appendix F** Advantage doubly robust estimator.

Supplementary Results: Table S1. Variance-covariance matrices of blip parameter estimates for the learned strategy based on multiple imputation analysis of one hundred data sets. **Figure S1.** Flow diagrams for the development set (A) and validation set (B). **Figure S2.** Comparison of recommendations from the original (A) or stringent (B) learned strategy and the RRT prescriptions received in the development set. **Figure S3.** Missing data patterns in the development set (A) and validation set (B).

Ethics approval and consent to participate

The MIMIC-III analysis received approval from the Institutional Review Boards of the Massachusetts Institute of Technology and Beth Israel Deaconess Medical Center (BIDMC). The AKIKI and AKIKI2 trials received approval from competent French legal authority (Comité de Protection des Personnes d'Ile de France VI, ID RCB 2013-A00765-40, NCT01932190 for AKIKI and ID RCB 2017-A02382-51, NCT03396757 for AKIKI 2) and consent of patient or relatives was obtained before inclusion.

Consent for publication

All authors have consented to the publication of the present manuscript, should the article be accepted by the Editor-in-chief upon completion of the refereeing process.

Availability of data and materials

The MIMIC-III data is publicly available at https://mimic.mit.edu. Anonymous participant data from the AKIKI trials is available under specific conditions. Proposals will be reviewed and approved by the sponsor, scientific committee, and staff on the basis of scientific merit and absence of competing interests. Once the proposal has been approved, data can be transferred through a secure online platform after the signing of a data access agreement and a confidentiality agreement.

Tables

Table 1 Baseline and evolving characteristics of the patients from the development set (MIMIC-III) and the validation set (AKIKI trials).

	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)

The crude strategy includes the following three decision rules that we derived from the blip parameter estimates $\hat{\psi}$. Decision rules are applicable to patients with severe AKI who have never initiated RRT at a given decision timepoint and whose previous recommendations from the crude strategy were never to initiate RRT (else, the crude strategy persist in its choice to initiate RRT). At the first decision timepoint (beginning of day 1), RRT should be initiated within 24 hours if the linear combination $-39.589 + 0.245 \times \text{age}_{t=1}\ (years) + 1.349 \times \text{creatinine}_{t=1}\ (mg/dL) + 3.409 \times \text{potassium}_{t=1}\ (mmoL/L)$ is positive. At the second decision timepoint (beginning of day 2), RRT should be initiated within 24 hours if $-7.747 + 0.514 \times \text{SOFA}_{t=2} + 0.095 \times \text{blood urea nitrogen}_{t=2}\ (mg/dL) - 63.874 \times |\text{pH}_{t=1} - \text{pH}_{t=2}| - 7.734 \times [\text{urine output}_{t=1} + \text{urine output}_{t=2}]$ is positive. At the third decision timepoint (beginning of day 3), RRT should be initiated within 24 hours if $5.397 - 19.316 \times \text{urine output}_{t=3}\ (ml/kg/h) + 1.922 \times [\text{blood urea nitrogen}_{t=3} / \text{blood urea nitrogen}_{t=1}]$ is positive. The $(-)_{t=1}$, $(-)_{t=2}$, $(-)_{t=3}$ subscripts refer to values measured just before the first, second, and third decision time point respectively (i.e., at the time of stage 3 KDIGO-AKI onset, stage 3 KDIGO-AKI + 24 hours, and stage 3 KDIGO-AKI + 48 hours respectively). AKI=Acute Kidney Injury. KDIGO=Kidney Disease Improving Global Outcomes. SOFA score=Sequential Organ Failure Assessment score.

[a] in the development set n=3748, in the validation set n=1068.

[b] in the development set n=3570, in the validation set n=869.

[c] in the development set n=3431, in the validation set n=718.

Table 3 Use of the learned strategies for individualized decision-making in three illustrative examples. In patient one (a man aged 58 years), disease severity (as described by SOFA score and arterial blood pH) and kidney function (as described by blood urea nitrogen, serum creatinine, and urine output) remain stable, and the crude strategy suggests against initiating RRT in the 72 hours following stage 3 KDIGO-AKI onset. In patient two (a woman aged 60 years), disease severity lessens over time, but kidney function deteriorates, and the crude strategy suggests initiating RRT on the third day following stage 3 KDIGO-AKI. In patient three (a woman aged 65 years), disease severity is stabilized 24 hours after stage 3 KDIGO-AKI onset, but kidney function has become critical, and the crude strategy suggests initiating RRT on the second day. Note that once a learned strategy recommends initiating RRT it persists in its recommendation until the third day regardless of patients' subsequent characteristics. AKI=Acute Kidney Injury. KDIGO=Kidney Disease Improving Global Outcomes. RRT=Renal Replacement Therapy. SOFA score=Sequential Organ Failure Assessment score.

	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

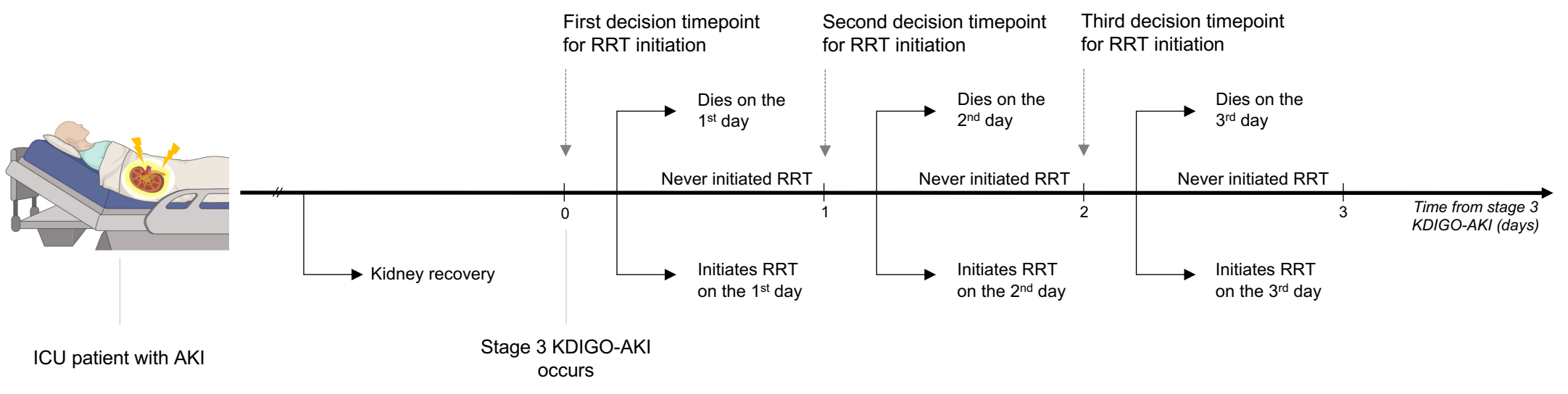
Figure legends

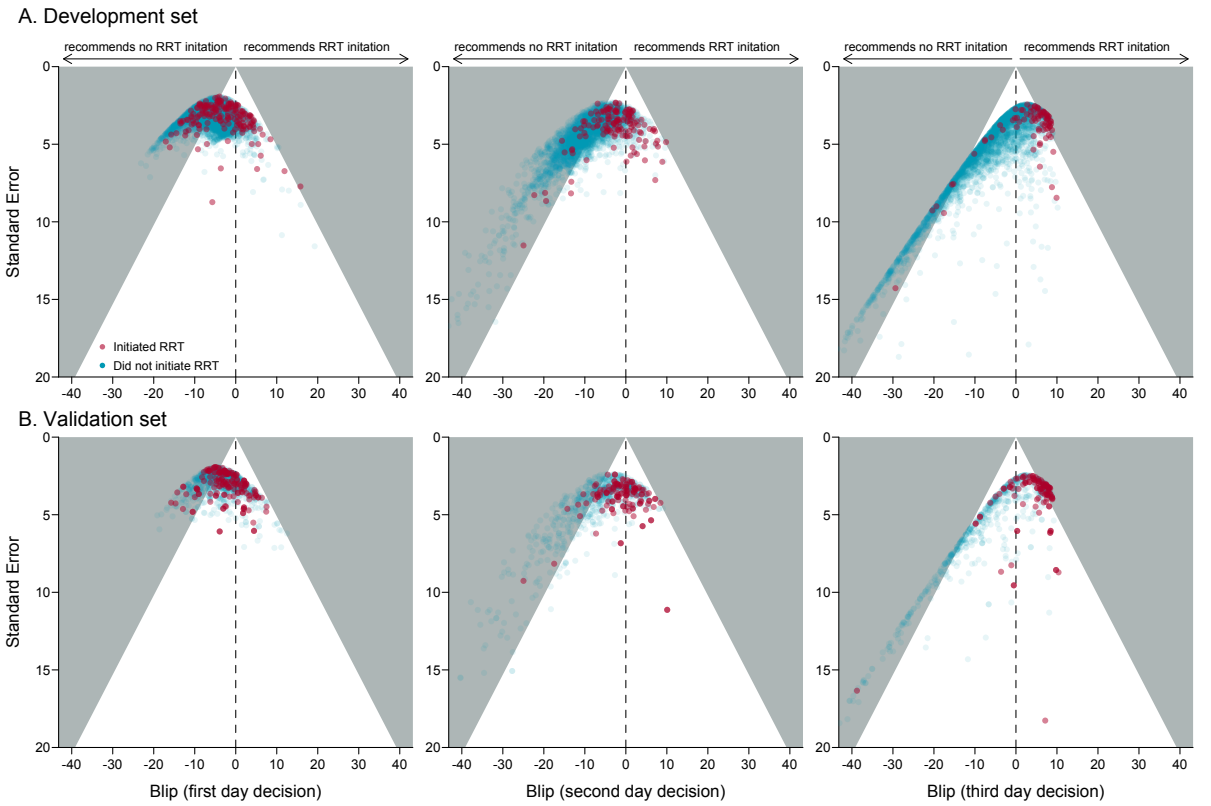
Figure 1 Possible trajectories of a single patient with acute kidney injury in our learning setup.
The first decision timepoint is defined as the time when stage 3 KDIGO-AKI occurs. In our setup, for a patient with stage 3 KDIGO-AKI, the decision rule to initiate RRT mimics that of clinicians i.e., decisions are re-evaluated every day—for three days in a row–given patients' evolving characteristics. Note that at a given decision timepoint a decision needs to be made only if a patient has neither initiated RRT nor died earlier. AKI=Acute Kidney Injury. ICU=Intensive Care Unit. KDIGO=Kidney Disease Improving Global Outcomes. RRT=Renal Replacement Therapy.

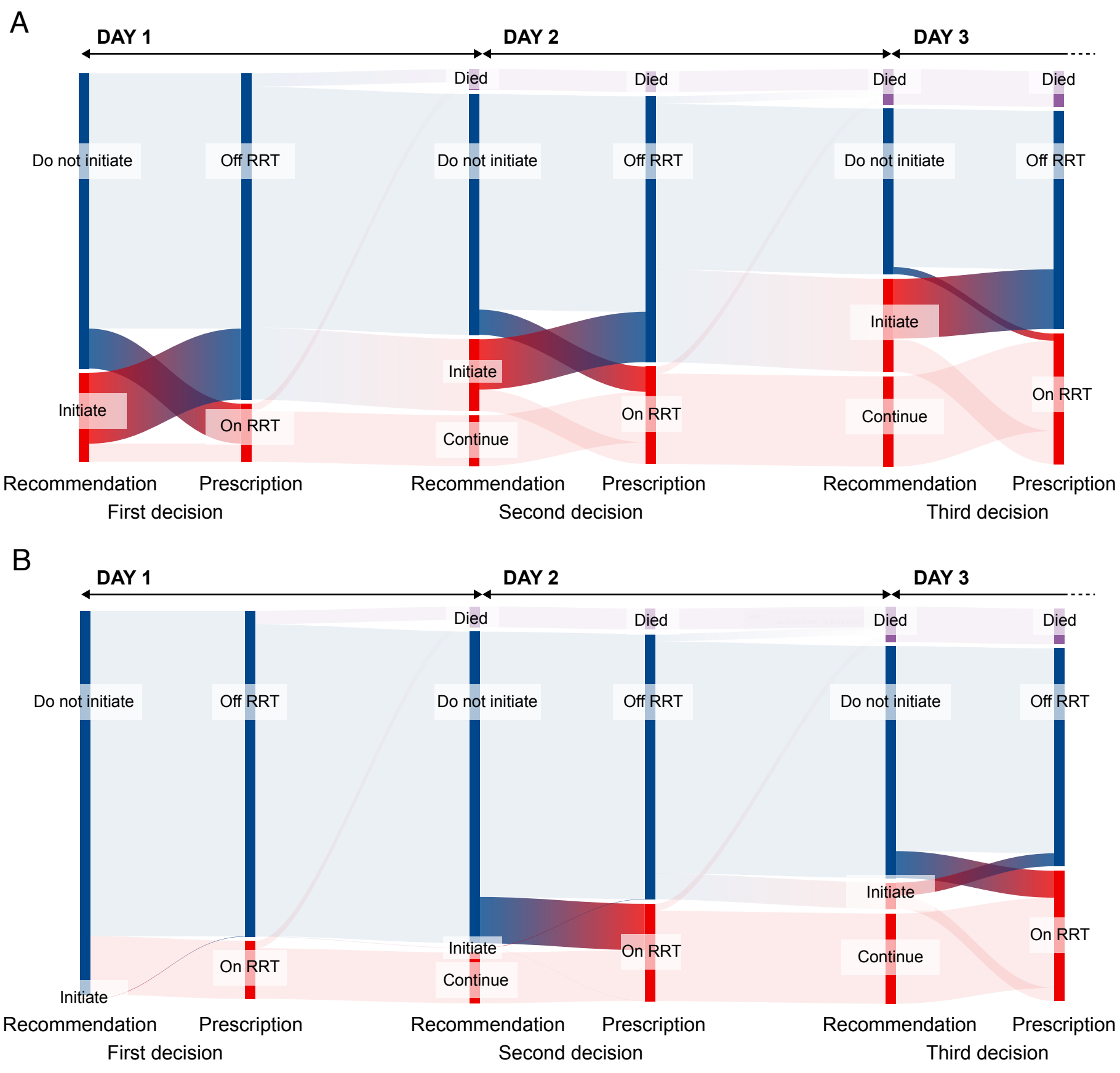
Figure 2 Recommendations from the learned strategies for patients in the development set (Panel A) and in the validation set (Panel B).
Each dot corresponds to a patient for whom a decision whether to initiate RRT needed to be made at the first (left-hand panels), second (middle panels), or third (right-hand panels) decision timepoint. Dot colors depict the RRT prescription observed for these patients. On the on *x*-axis, predicted blips indicate on a HFD60 scale the magnitude of individual-patient harm (negative blips) or benefit (positive blips) from initiating RRT at a particular timepoint. Vertical dashed lines indicate no effect. Uncertainty in the individual-patient blips is represented on *y*-axis. Dots falling in grey-shaded aeras represent patients for whom there is evidence of either harm (left-hand aeras), or benefit (right-hand aeras) from RRT initiation at the 0.05 alpha level. The crude strategy would recommend initiating RRT at a given timepoint if a patient's dot fell on the right-hand side of the dashed line. On the other hand, the stringent strategy would recommend initiating RRT at a given timepoint only if a patient's dot fell in the right-hand grey-shaded aera.

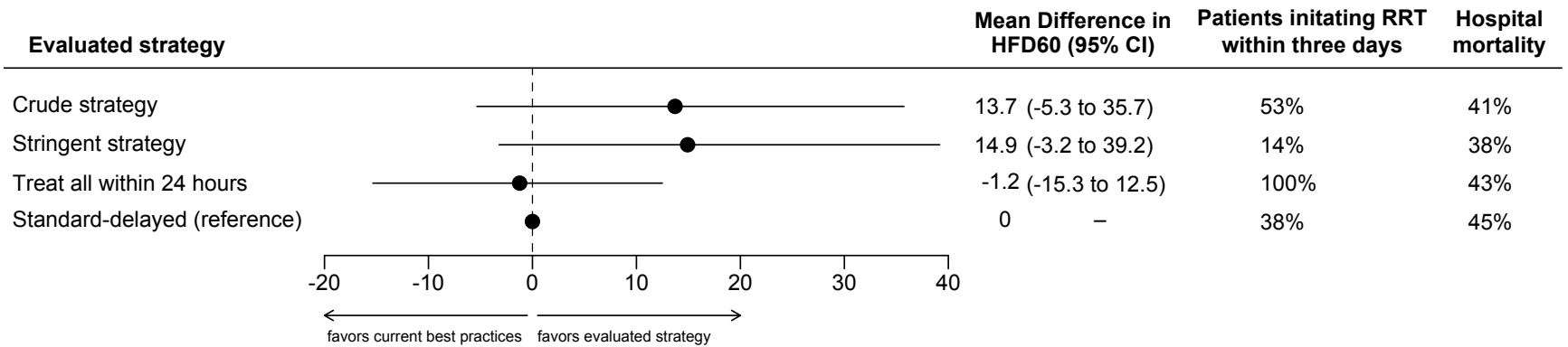
Figure 3 Comparison of recommendations from the crude (Panel A) or stringent (Panel B) strategy and the prescriptions received in the validation set.
Prescriptions received are denoted 'On RRT' or 'Off RRT.' The bar heights represent the proportions of patients in each category. At each decision timepoint, recommendation and prescription of RRT appear in red while the absence of recommendation or prescription of RRT is shown in blue. Discrepancies between recommendations and prescriptions are shown in brighter colors. Note that when patients initiated RRT (sometimes in contradiction with the strategy's recommendation) the strategy never recommends stopping it afterward.

Figure 4 External validation of the learned strategies' benefit as compared to current best practices (i.e., a standard-delayed strategy).
The mean difference in HFD60 represent the causal effects of implementing a strategy compared to current best practices. The "crude strategy" refers to the strategy derived from the blip parameter estimates given in Table 2. The "stringent strategy" refers to a strategy that recommends initiating RRT only when there is evidence at the 0.05 threshold that a patient will benefit from RRT initiation. The "treat all within 24 hours strategy" designates a strategy to initiate RRT in all patients within 24 hours regardless of emergency criteria. HFD60=Hospital-Free Days at day 60. CI=Confidence Interval. RRT=Renal Replacement Therapy.


First decision timepoint for RRT initiation
Second decision timepoint for RRT initiation
Third decision timepoint for RRT initiation
Dies on the 1st day
Dies on the 2nd day
Dies on the 3rd day
Never initiated RRT
Never initiated RRT
Never initiated RRT
0
1
2
3
Time from stage 3 KDIGO-AKI (days)
Kidney recovery
Initiates RRT on the 1st day
Initiates RRT on the 2nd day
Initiates RRT on the 3rd day
ICU patient with AKI
Stage 3 KDIGO-AKI occurs


A. Development set
recommends no RRT initation
recommends RRT initation
Standard Error
Initiated RRT
Did not initiate RRT
B. Validation set
Blip (first day decision)
Blip (second day decision)
Blip (third day decision)


A
DAY 1
DAY 2
DAY 3
Died
Do not initiate
Off RRT
Initiate
On RRT
Continue
Recommendation
Prescription
First decision
Second decision
Third decision
B


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%
-20
-10
0
10
20
30
40
favors current best practices
favors evaluated strategy

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,[1] François PETIT*,[2] Stéphane GAUDRY*,[3] Élise DIARD,[4] Jean-Pierre QUENOT,[5] Didier DREYFUSS,[6] Viet-Thi TRAN,[7] Raphaël PORCHER.[8]

* These authors contributed equally as second authors.

ADITIONAL FILE

Table of Contents

Supplementary Methods

Setup notations

- t : decision timepoint, $t \in \{1,2,3\}$.
- A_t : treatment observed at time t.
- H_t : history of variables collected up to time t including the treatment received before time t but excluding the treatment received at time t.
- H_t^f : subset of H_t relevant for prognosis.
- H_t^γ : subset of H_t relevant for the effect of treatment initiation at time t.
- Y : outcome of interest.
- $Y^{\bar{a}_t, \underline{a}_{t+1}^{opt}}$: potential outcome corresponding to the outcome that would have been observed if treatments $a_1, \dots, a_t$ had been delivered at decision timepoints $1, \dots, t$ and (possibly contrary to fact) all subsequent treatment decisions had been optimal. We use over and underline notations to indicate the past and future treatments respectively i.e., $\bar{a}_t = (a_1, \dots, a_t)$ and $\underline{a}_t = (a_t, \dots, a_3)$.

Summary of notations introduced in the appendix

- $e_t(H_t) = \mathbb{E}[A_t|H_t]$: propensity score at time t.
- $f_t(h_t) = \mathbb{E}\left[Y^{\bar{a}_{t-1}, 0, \underline{a}_{t+1}^{opt}} | H_t = h_t\right]$: treatment-free function at time t.
- $\gamma_t(a_t, h_t) = \mathbb{E}\left[Y^{\bar{a}_{t-1}, a_t, \underline{a}_{t+1}^{opt}} - Y^{\bar{a}_{t-1}, 0, \underline{a}_{t+1}^{opt}} | H_t = h_t\right]$: blip function at time t.
- $\widetilde{Y}_t = \widehat{\mathbb{E}}[Y^{\bar{a}_t, \underline{a}_{t+1}^{opt}} | H_t, A_t]$: pseudo-outcomes at time t.
- $\widetilde{w}_t(H_t) = |A_t - \hat{e}_t(H_t)|$: overlap weights at time t.
- $\tau = (H_3, A_3, Y)$: the observable full trajectory of a patients.
- $\mathcal{T}$: the space of trajectories.
- $\mathcal{H}_t$: the space of histories of variables collected up to time t.
- $\pi = (\pi_1, \pi_2, \pi_3)$ with $\pi_t: \mathcal{H}_t \longrightarrow \{0,1\}$ for $t = 1,2,3$: the non-stationary deterministic policy* π.
- $H_t^{(i)} = \Phi$: indicates that at time t, patient i is in a terminal state (i.e., death).

* Note that throughout the paper, we use the term strategy rather than policy. For the remainder of this appendix, these can be taken to be synonymous.

Doubly robust dynamic treatment regimen via weighted least squares

The procedure (termed dWOLS) was formally introduced and described in detail by Wallace and Moodie.[1] Succinctly, the method requires that for each decision timepoint $t = 1,2,3$, we posit models for the propensity scores $e_t(H_t) = \mathbb{E}[A_t|H_t]$ as well as the treatment-free $f_t(\cdot)$, and blip $\gamma_t(\cdot)$ functions. Treatment-free and blip functions are defined as $f_t(h_t) = \mathbb{E}\left[Y^{\bar{a}_{t-1},0,\underline{a}_{t+1}^{opt}}|H_t = h_t\right]$, and $\gamma_t(a_t, h_t) = \mathbb{E}\left[Y^{\bar{a}_{t-1},a_t,\underline{a}_{t+1}^{opt}} - Y^{\bar{a}_{t-1},0,\underline{a}_{t+1}^{opt}}|H_t = h_t\right]$ so that $f_t(h_t) + \gamma_t(a_t, h_t) = \mathbb{E}\left[Y^{\bar{a}_t,\underline{a}_{t+1}^{opt}}|H_t = h_t, A_t = a_t\right]$.[†] The estimation of $f_t(\cdot)$ and $\gamma_t(\cdot)$ starts at $t = 3$ by regressing $Y^{\bar{a}_3,\underline{a}_{3+1}^{opt}} = Y$ onto $(H_3^f\ ,\ A_3H_3^\gamma)$ via weighted least squares with weights $\widetilde{w}_3(H_3) = |A_3 - \hat{e}_3(H_3)|$. The procedure then follows a backward stepwise approach where we substitute all unobserved potential outcomes by pseudo-outcomes. Specifically, for $t = 2,1$, we build pseudo-outcomes $\widetilde{Y}_t = \widehat{\mathbb{E}}[Y^{\bar{a}_t,\underline{a}_{t+1}^{opt}}|H_t, A_t]$ by taking naive outcomes Y and summing up subsequent regrets i.e., $\widetilde{Y}_t = Y + \sum_{k=t+1}^{3} \max \hat{\gamma}_k(1, H_t), 0 - \hat{\gamma}_k(A_t, H_t)$. Pseudo-outcomes at time t represent the outcomes that would have been observed if treatment decisions had been optimal from time $t+1$ onwards. We then regress $\widetilde{Y}_t$ onto $(H_t^f\ ,\ A_tH_t^\gamma)$ via weighted least squares with weights $\widetilde{w}_t(H_t) = |A_t - \hat{e}_t(H_t)|$. Using these overlap weights provide double robustness and enhance sample efficiency. Note that the dWOLS estimation procedure does not require making a Markovian assumption. It only requires assuming that for each decision timepoints either the variables causing renal replacement therapy (RRT) initiation, or the prognosis variables were measured. Because we considered that once initiated, RRT is not stopped in the three days following stage 3 KDIGO-AKI, analysis for each decision timepoint was limited to those participants who had not initiated RRT until this decision timepoint (as those who had initiated RRT had no treatment decision to make).

Variable selection

For each decision timepoint, the variables we considered for modeling the probability of RRT initiation withing 24 hours were: blood urea nitrogen, serum potassium, arterial blood pH, and urine output. For each decision timepoint, the variables we considered for predicting hospital-free days at day 60 (HFD60) were: age, weight, gender, SOFA score, serum creatinine, blood urea nitrogen, serum potassium, arterial blood pH, and urine output. We considered the evolving values of the aforementioned variables prior to the decision timepoint of interest. The same variables were considered in development and validation analyzes.

[†] This last equality uses the sequential ignorability assumption i.e., $Y^{\bar{a}_{t-1},\underline{a}_t} \perp\!\!\!\perp A_t|H_t$ for all t.

Missing data management

Missing data were handled through multiple imputations by chained equations using outcomes as well as all aforementioned predictors in the imputation models. One hundred independent imputed data sets were generated and analyzed separately. Variance-covariance matrices of blip functions parameters were estimated using the bootstrap (999 iterations). Estimates were then pooled using Rubin's rules.

Importance sampling for policy evaluation

We used importance sampling for policy evaluation in reinforcement learning[2] to estimate hospital mortality as well as the proportion of patients who would initiate RRT within three days under a learned strategy. The approach is similar to inverse propensity weighting as used in the context of marginal structural modeling in epidemiology.[3] Succinctly, denoting $\tau = (H_3, A_3, Y) \in \mathcal{T}$ the observable full trajectory of a patient and $R: \mathcal{T} \longrightarrow \mathbb{R}$ any reward function of the trajectory, the expected reward under a different strategy, say the non-stationary deterministic strategy π i.e., $\pi = (\pi_1, \pi_2, \pi_3)$ with $\pi_t: \mathcal{H}_t \longrightarrow \{0,1\}$ for $t = 1,2,3$, can be estimated by

$$\widehat{\mathbb{E}}_{\tau\sim\pi}[R(\tau)] = n^{-1}\sum_{i=1}^{n} R\left(\tau^{(i)}\right)\prod_{k=1}^{3}\frac{\mathbb{I}\left[\pi_k\left(H_k^{(i)}\right) = A_k^{(i)}\right]}{\hat{e}_k(H_k^{(i)})^{A_k^{(i)}}\{1-\hat{e}_k(H_k^{(i)})\}^{1-A_k^{(i)}}}. \qquad (1)$$

To estimate hospital mortality, the reward function we used was $R\left(\tau^{(i)}\right) = 1$ if patient i died in the hospital and $R\left(\tau^{(i)}\right) = 0$ otherwise. To estimate the proportion of patients who would initiate RRT within three days under our learned strategies, we used the reward function $R\left(\tau^{(i)}\right) = 1 - \prod_{k=1}^{3}\mathbb{I}\left[A_k^{(i)} = 0\right]$ which outputs one whenever patient i initiated RRT at any time along their observed trajectory. The estimator above straightforwardly handles the patients who died before day 3, provided we consider that the strategy π stops prescribing treatment once a patient has died i.e., $\pi_k(\Phi) = 0$, and that no RRT was prescribed to the patients who have died i.e., $e_k(\Phi) = 0$.[‡] The variables we considered for modeling the propensity scores are identical to those given in the previous section. To improve efficiency, we used the weighted version of the estimator above that is given in equation 3 from Precup et al.[2]

[‡] For the sake of clarity, we denoted $H_t^{(i)} = \Phi$ when patient i is in a terminal state (i.e., death) at time t.

Advantage doubly robust estimator

Although the estimator given in equation 1 accounts for the patients who died before day 3, it only uses trajectories that match the policy π exactly, which can make policy evaluation sample inefficient. To estimate the causal effect of implementing the original, stringent, or treat all strategies compared to current best practices, we used the cross-fitted advantage doubly robust (ADR) estimator with terminal state for strategy evaluation given in the Algorithm 2 from Nie et al.[4] The ADR estimator allows the evaluation of when-to-treat-policies exploiting the subparts of trajectories that match the policy π. The original, stringent, and treat all strategies are all regular when-to-treat policies in the sense of Definition 1b from Nie et al. The ADR estimator is more data efficient but also more robust than the estimator given in equation (1). Briefly, this estimator relies on the decomposition into a sum of local advantages of the relative value of any given strategy in comparison to that of the never-treating policy $\mathbf{0}$, following Lemma 1 of Murphy.[5] The ADR estimand is $\Delta(\pi, \mathbf{0}) = \mathbb{E}_{\tau\sim\pi}(Y) - \mathbb{E}_{\tau\sim\mathbf{0}}(Y)$ where, π denotes the strategy to be tested, and zero is the never-treating policy. The causal effects of implementing the original, stringent, or treat all strategies compared to a "best practices policy" denoted π_{bp}, are given by

$$\widehat{\Delta}(\pi, \pi_{bp}) = \widehat{\Delta}(\pi, \mathbf{0}) \; - \; \widehat{\Delta}(\pi_{bp}, \mathbf{0}).$$

As in the dWOLS procedure, the ADR estimator does not need any structural (e.g., Markovian) assumptions. As Nie et al.,[4] we estimated all the nuisance components using cross-fitting to reduce the effect of own-observation bias.

For each decision timepoint, the variables we considered for modeling the probability of RRT initiation withing 24 hours were: blood urea nitrogen, serum potassium, arterial blood pH, and urine output. For each decision timepoint the variables we considered for predicting hospital-free days at day 60 (HFD60) were: age, weight, gender, SOFA score, serum creatinine, blood urea nitrogen, serum potassium, arterial blood pH, and urine output.

Missing data were handled through multiple imputations by chained equations using outcomes as well as all aforementioned predictors in the imputation models. Twenty independent imputed data sets were generated and analyzed separately. The variances of the estimators were estimated using the bootstrap (999 iterations). Estimates were then pooled using Rubin's rules.

Supplementary Results

Table S1. Variance-covariance matrices of blip parameter estimates for the learned strategy based on multiple imputation analysis of one hundred data sets.

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

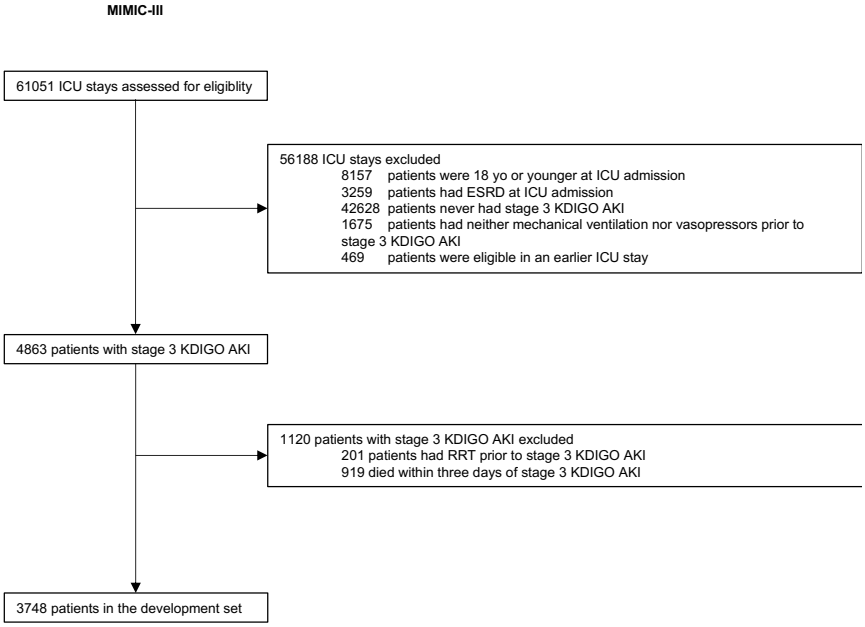
$$\hat{\psi}_t{}^T H_t \pm 1.96 \times \sqrt{H_t{}^T \widehat{M}_t H_t}.$$

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		

Units are years for age; mg/dL for creatinine; mmol/L for potassium; mg/dL for blood urea nitrogen; ml/kg/h for urine output. The $(-)_{t=1}$, $(-)_{t=2}$, $(-)_{t=3}$ subscripts refer to values measured just before the first, second, and third decision time point respectively.

Figure S1. Flow diagrams for the development set (A) and validation set (B).

A



B

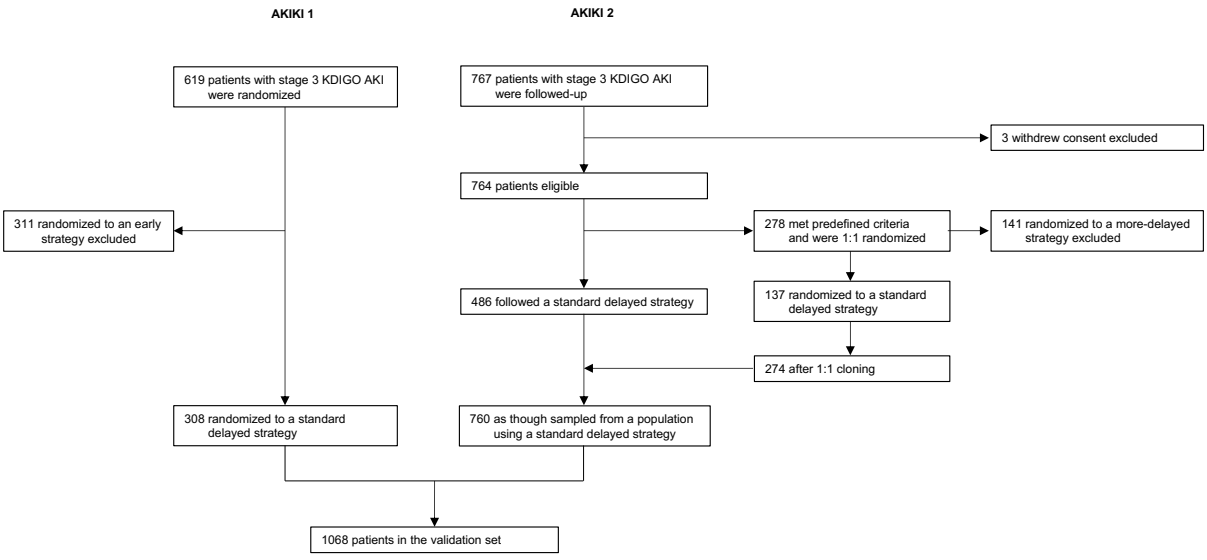
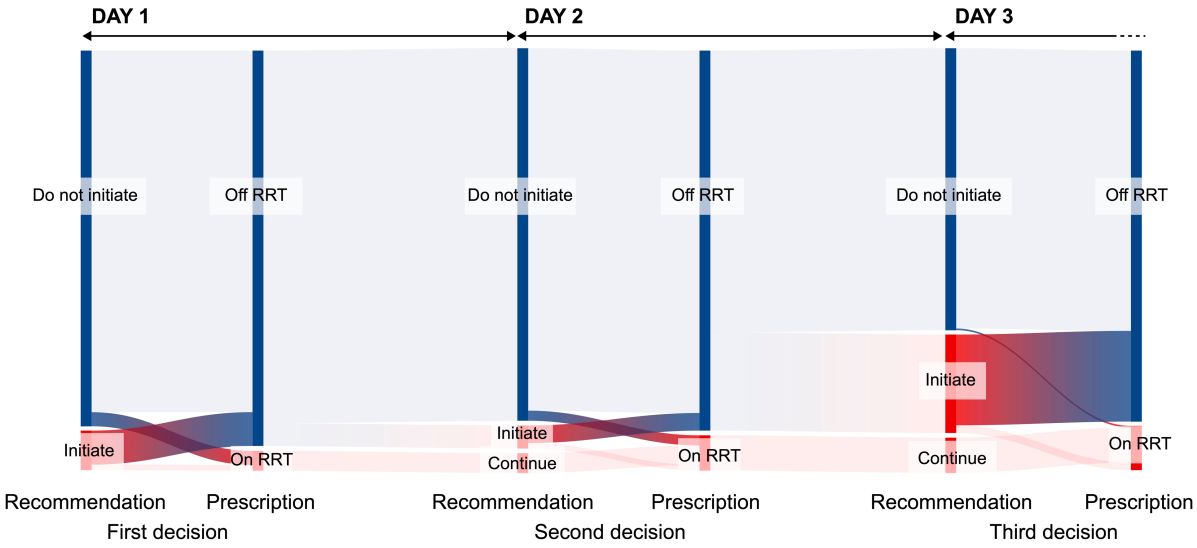


Figure S2. Comparison of recommendations from the original (A) or stringent (B) learned strategy and the RRT prescriptions received in the development set.

A



B

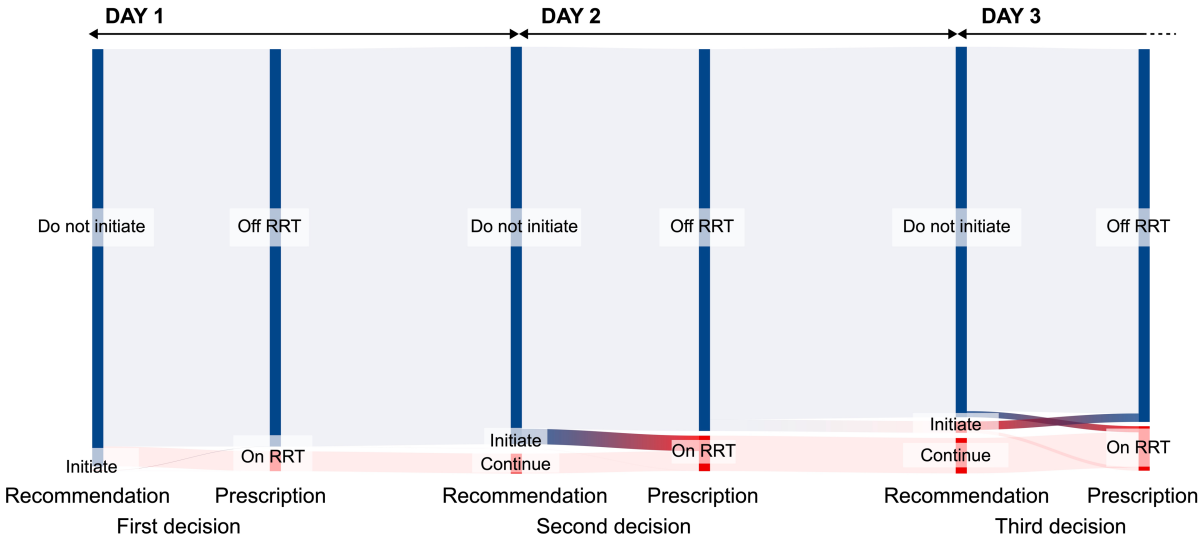
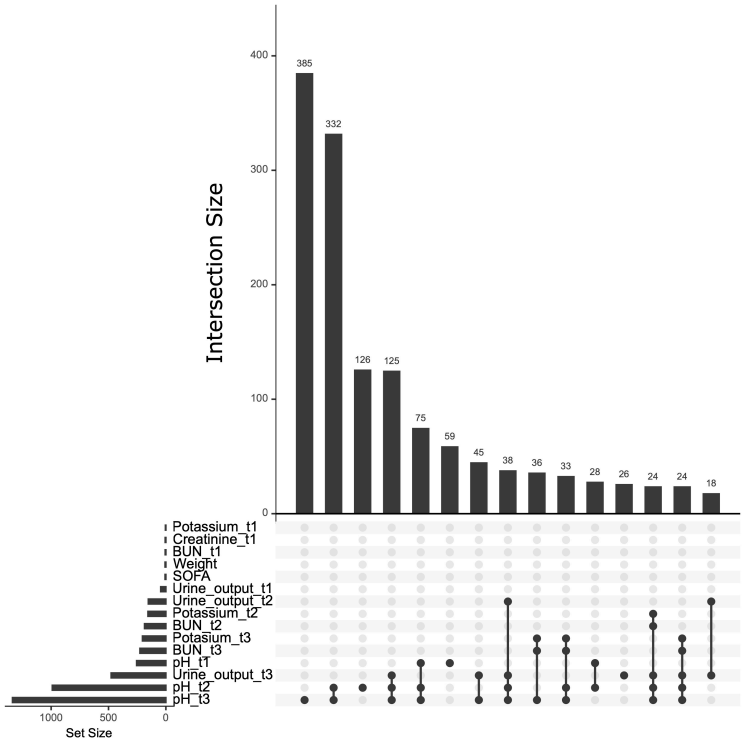
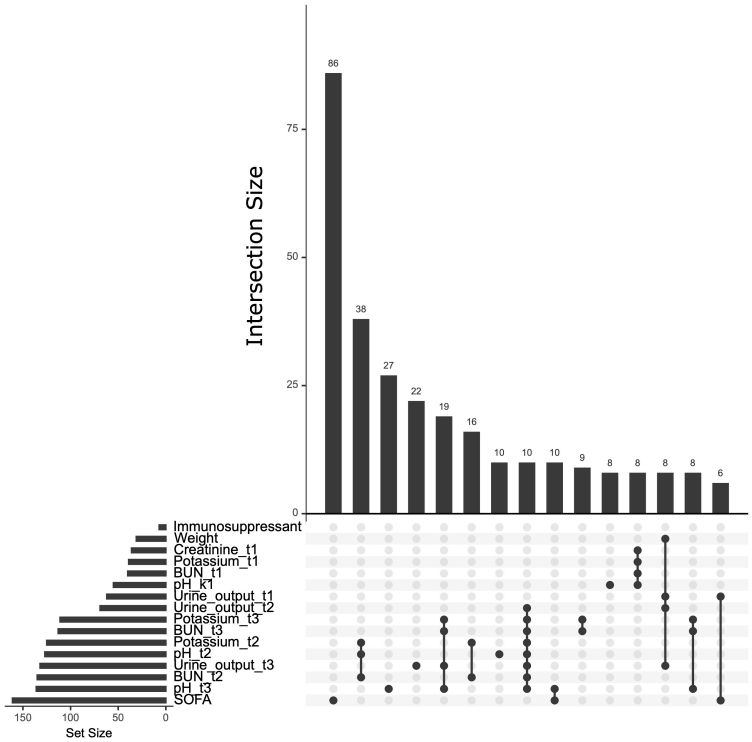


Figure S3. Missing data patterns in the development set (A) and validation set (B).

A



B



References

1 Wallace MP, Moodie EEM. Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics* 2015; **71**: 636–44.

2 Precup D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series* 2000; p. 80.

3 Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**: 550–60.

4 Nie X, Brunskill E, Wager S. Learning when-to-treat policies. *Journal of the American Statistical Association* 2021; **116**: 392–409.

5 Murphy SA. A Generalization Error for Q-Learning. *J Mach Learn Res* 2005; **6**: 1073–97.