Supplemental Material

Sensitivity Analysis

To ensure that our algorithms were not fitting spurious data, we created 100 "pseudo-interventions"; entirely post hoc treatment assignments that could have no true bearing on outcome in any subset of patients as it is not associated with any treatment. The uplift algorithms, as expected, failed to identify patients that would benefit from the pseudo-interventions. The distribution of T-scores associated with the pseudointerventions are given in supplemental Table 3 and, as expected, have a mean close to 0 and a standard deviation near 1. Supplemental Table 3 also illustrates how many trials out of the 100 pseudointerventions would result in a spurious finding of statistical significance. We found this to be the case in no more than 2 of the 100 pseudointerventions using any of the algorithms.

To ensure that our algorithm wasn't spuriously fitting on the test set, we performed a 100-fold Monte-Carlo cross-validation with the entire dataset. For the Z-Learner, the mean t-score across the 100 splits was -1.77 ± 0.93, for the T-Learner -1.73 ± 0.86 and for the X-Learner -1.93 ± 1.07. The prognostic target, as expected, had poorer performance as an uplift score in cross validation with t-score -1.18 ± 1.21. In contrast, a purely random uplift model mirrored the expected null distribution with t-statistic -0.08 ± 1.02. Performance distributions are presented in Supplemental Figure 1.

Finally, uplift performance did not differ substantially when the 8.5% of patients without a post-alert creatinine available were excluded from the analysis, as seen below:

Method	Interaction P All patients	Interaction P Patients with follow-up
Z-method	0.013	0.016
X-method	0.025	0.025
T-method	0.024	0.024
Prognostic method	0.299	0.16

Supplemental Methods

Uplift Modeling

Our first approach uses a two-equation prognostic model, referred to as the "T-Learner". Here, we create two independent multivariable linear models, one for the alert group and the other for the control group, examining factors associated with the primary outcome. These two models are then applied to test set patients, and the difference in predicted outcome is our estimate of the marginal benefit of treatment (i.e. uplift).[1-4]

In our second approach, we created a model, which we name a "Z-Learner", employing a transformation to the primary outcome: we reset the outcomes in the training set to 1 * the outcome in the alert group and -1 * the outcome in the usual care group. We then trained a single linear regression model on this transformed outcome. The predictions from that model serve as the marginal benefit of treatment in the test set (i.e. uplift).[5]

In our third approach, we used a cross-prediction approach "X-Learner".[6] Here, a prognostic model is built using data from usual care patients, and predictions are made on alert patients. A second model, trained on alert patients, is applied to usual care patients. The residuals of these models (the difference between observed outcome and expected outcome) is due to both error and the marginal effect of the treatment. We train a third model on these residuals, which is finally applied to test set data to predict the marginal benefit of treatment (i.e. uplift).

We used a similar approach as above for the creation of the prognostic model. The prognostic model was a linear regression model, trained to predict the primary outcome, using all training patients (irrespective of treatment status).

As all models used linear regression, or combinations of linear regressions as the base function, we were able to maintain a uniform approach to modeling strategy between the uplift models and including the prognostic model. Within the training set, we used 100-fold Monte-Carlo cross-validation to train 100 models in 71.5% of the training set, and validate those models in the remaining 28.5% of the training set. We elected to use an ensemble of models in order to minimize variance; uplift estimates in the test set were the arithmetic mean of the 100 training set models. For all regressions, a base set of clinical predictors were always included (see below). From this point, forward feature selection was performed based on model performance in the validation set. From the perspective of the totality of data available, then, we trained models on 50% of the data, validated them on 20%, and tested them on the remaining 30%.

Covariate Selection

Electronic health records have a wealth of covariates. To select those that may be important for modeling marginal treatment effect, we used Monte-Carlo cross-validation and a forward-selection procedure to select variables that would be used in modeling. All modeling used the same base set of covariates, chosen based upon prior literature examining prognostic factors in AKI and expert opinion.[7-11] These variables included baseline: age, sex, race, surgical vs. medical admission, intensive care unit location, serum creatinine, serum bicarbonate, blood urea nitrogen, hemoglobin, white blood cell count, platelet count, serum potassium and serum sodium. Candidate predictors that could be chosen via forward selection appear in supplemental Table 1. We elected not to include ICD-9 based comorbidity data as candidate predictors as these may be diagnosed after randomization, biasing results.

Supplemental References

1. Rubin DB. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics.* 1978:34-58.
2. Lo VS. The true lift model: a novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter.* 2002;4(2):78-86.
3. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: A review. *The review of Economics and Statistics.* 2004;86(1):4-29.
4. Rubin DB. Assignment to Treatment Group on the Basis of a Covariate. *Journal of educational Statistics.* 1977;2(1):1-26.
5. Jaskowski M, Jaroszewicz S. Uplift modeling for clinical trial data. Paper presented at: ICML Workshop on Clinical Data Analysis2012.
6. Künzel S, Sekhon J, Bickel P, Yu B. Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning. *arXiv preprint arXiv:170603461.* 2017.
7. Bagur R, Webb JG, Nietlispach F, et al. Acute kidney injury following transcatheter aortic valve implantation: predictive factors, prognostic value, and comparison with surgical aortic valve replacement. *Eur Heart J.* 2010;31(7):865-874.
8. Brown JR, Kramer RS, MacKenzie TA, Coca SG, Sint K, Parikh CR. Determinants of acute kidney injury duration after cardiac surgery: an externally validated tool. *Ann Thorac Surg.* 2012;93(2):570-576.
9. Bihorac A, Delano MJ, Schold JD, et al. Incidence, clinical predictors, genomics, and outcome of acute kidney injury among trauma patients. *Ann Surg.* 2010;252(1):158-165.
10. Grams ME, Sang Y, Ballew SH, et al. A Meta-analysis of the Association of Estimated GFR, Albuminuria, Age, Race, and Sex With Acute Kidney Injury. *Am J Kidney Dis.* 2015;66(4):591-601.
11. Koyner JL, Adhikari R, Edelson DP, Churpek MM. Development of a Multicenter Ward-Based AKI Prediction Model. *Clin J Am Soc Nephrol.* 2016;11(11):1935-1943.

Supplemental Table 1: Eligible Features Considered for Modeling in a Randomized Trial of Acute Kidney Injury Alerts

Base Features (Included in Every Model)	**Percent Missing**	**Features Eligible for Inclusion via Forward Selection**	**Percent Missing**
Demographics		*Labs:*	
Age	0.7	Alkaline phosphatase	28.2
Sex	0.7	Basophil count	15.9
Race	0.0	Chloride	1.8
Surgical admission	0.0	Creatinine slope	0.0
Intensive Care status	0.0	Eosinophil percentage	15.9
		Glucose	0.8
Labs:		Hematocrit	0.29
Serum creatinine	0.0	Lactate	39.1
Blood urea nitrogen	0.17	Lymphocyte count	15.9
Hemoglobin	0.25	Magnesium	10.4
Platelet count	0.33	Mean corpuscular hemoglobin	0.33
Serum bicarbonate	0.17	Monocyte count	15.88
Serum potassium	0.50	Neutrophil count	18.68
Serum sodium	0.17	Phosphorus	23.9
White blood cell count	0.33	Prothrombin time	11.5
		Red cell distribution width	0.33
		Total bilirubin	28.2
		Urinalysis protein	49.6
		Urinalysis specific gravity	49.6

Supplemental Table 2A: Characteristics Predictive of Alert Harm or Benefit (T-Method) among Patients in a Randomized Trial of Acute Kidney Injury Alerts

	Unlikely to Benefit from Alert (n=258)	Likely to Benefit from Alert (n=425)	P-value
Demographics			
Age, yrs	57 (18)	63 (16)	<0.001
Male gender	180 (71)	196 (46)	<0.001
Black	59 (23)	138 (33)	0.007
In ICU at Randomization	74 (29)	119 (28)	0.85
Surgical Admission	109 (42)	157 (37)	0.17
Comorbidities			
Cerebrovascular Disease	34 (13)	61 (14)	0.67
Chronic Kidney Disease	94 (36)	94 (22)	<0.001
Congestive Heart Failure	99 (38)	132 (31)	0.05
Diabetes	83 (32)	126 (30)	0.49
Liver Disease	45 (17)	57 (13)	0.15
Malignancy	45 (17)	113 (27)	0.006
Metastatic Disease	13 (5)	41 (10)	0.03
Laboratory Values			
Baseline Creatinine, mg/dL	1.19 (0.79 - 1.74)	0.77 (0.51 - 1.11)	<0.001
Alert Creatinine, mg/dL	1.57 (1.19 - 2.15)	1.18 (0.83 - 1.67)	<0.001
Hemoglobin, g/dL	10.5 (9.2 - 12.1)	10.2 (8.9 - 11.6)	0.07
Phosphorus, mg/dL	4.2 (3.4 - 5.1)	3.5 (2.9 - 4.2)	<0.001
Potassium, meq/L	4.3 (3.9 - 4.7)	4.2 (3.8 - 4.5)	0.007
Sodium, meq/L	138 (135 - 141)	137 (134 - 140)	0.009
Severity of Illness			
SOFA Score	2 (1 – 5)	2 (1 – 4)	0.08
Medication Exposures			
NSAIDs	16 (6)	30 (7)	0.92
Vasopressors	48 (19)	75 (18)	0.75
Loop Diuretic	97 (38)	147 (35)	0.43
ACE-Inhibitor / Angiotensin Receptor Blocker	43 (17)	94 (22)	0.08
Antibiotics	152 (59)	285 (67)	0.03
Intravenous Contrast	44 (17)	101 (24)	0.04
Randomization Status			
Alert	133 (52)	217 (51)	0.90

Characteristics at randomization of patients in the test set predicted to be unlikely to benefit from alert vs. likely to benefit from alert based on T-method score <=0 or >0 respectively. Note that comorbidity codes and medication exposures are not used in model-building, and are produced here to provide a clearer clinical phenotype for readers.

Supplemental Table 2B: Characteristics Predictive of Alert Harm or Benefit (X-Method) among Patients in a Randomized Trial of Acute Kidney Injury Alerts

	Unlikely to Benefit from Alert (n=264)	Likely to Benefit from Alert (n=419)	P-value
Demographics			
Age, yrs	58 (18)	63 (16)	<0.001
Male gender	184 (71)	192 (46)	<0.001
Black	59 (22)	138 (33)	0.003
In ICU at Randomization	74 (28)	119 (28)	0.92
Surgical Admission	112 (42)	154 (37)	0.14
Comorbidities			
Cerebrovascular Disease	35 (13)	60 (14)	0.70
Chronic Kidney Disease	98 (37)	90 (22)	<0.001
Congestive Heart Failure	100 (38)	131 (31)	0.08
Diabetes	86 (33)	123 (29)	0.37
Liver Disease	44 (17)	58 (14)	0.31
Malignancy	45 (17)	113 (27)	0.003
Metastatic Disease	14 (5)	40 (10)	0.05
Laboratory Values			
Baseline Creatinine, mg/dL	1.20 (0.81 - 1.73)	0.77 (0.51 - 1.10)	<0.001
Alert Creatinine, mg/dL	1.58 (1.20 - 2.16)	1.17 (0.82 - 1.66)	<0.001
Hemoglobin, g/dL	10.4 (9.2 - 12.0)	10.2 (8.9 - 11.6)	0.12
Phosphorus, mg/dL	4.2 (3.4 - 5.0)	3.5 (2.9 - 4.2)	<0.001
Potassium, meq/L	4.3 (3.9 - 4.7)	4.2 (3.8 - 4.5)	0.004
Sodium, meq/L	138 (135 - 141)	137 (134 - 140)	0.008
Severity of Illness			
SOFA Score	2 (1 – 5)	2 (1 – 4)	0.56
Medication Exposures			
NSAIDs	16 (6)	30 (7)	0.58
Vasopressors	50 (19)	73 (17)	0.62
Loop Diuretic	99 (38)	145 (35)	0.44
ACE-Inhibitor / Angiotensin Receptor Blocker	43 (16)	94 (22)	0.051
Antibiotics	157 (60)	280 (67)	0.051
Intravenous Contrast	47 (18)	98 (23)	0.08
Randomization Status			
Alert	136 (52)	214 (51)	0.91

Characteristics at randomization of patients in the test set predicted to be unlikely to benefit from alert vs. likely to benefit from alert based on X-method score <=0 or >0 respectively. Note that comorbidity codes and medication exposures are not used in model-building, and are produced here to provide a clearer clinical phenotype for readers.

Supplemental Table 2C: Characteristics Predictive of Alert Harm or Benefit (Prog-Method) among Patients in a Randomized Trial of Acute Kidney Injury Alerts

	Unlikely to Benefit from Alert (n=341)	Likely to Benefit from Alert (n=342)	P-value
Demographics			
Age, yrs	64 (15)	58 (18)	<0.001
Male gender	204 (58)	172 (52)	0.07
Black	109 (31)	88 (26)	0.13
In ICU at Randomization	11 (3)	182 (54)	<0.001
Surgical Admission	116 (33)	150 (45)	0.003
Comorbidities			
Cerebrovascular Disease	46 (13)	49 (15)	0.62
Chronic Kidney Disease	118 (34)	70 (21)	<0.001
Congestive Heart Failure	122 (35)	109 (32)	0.45
Diabetes	115 (33)	94 (28)	0.14
Liver Disease	51 (15)	51 (15)	0.86
Malignancy	80 (23)	78 (23)	0.96
Metastatic Disease	31 (9)	23 (7)	0.31
Laboratory Values			
Baseline Creatinine, mg/dL	0.98 (0.73 - 1.52)	0.75 (0.50 - 1.15)	<0.001
Alert Creatinine, mg/dL	1.49 (1.10 - 1.98)	1.19 (0.82 - 1.70)	<0.001
Hemoglobin, g/dL	11.1 (9.9 - 12.6)	9.5 (8.5 - 10.9)	<0.001
Phosphorus, mg/dL	3.7 (3.0 - 4.4)	3.7 (3.0 - 4.5)	0.78
Potassium, meq/L	4.2 (3.9 - 4.6)	4.2 (3.8 - 4.5)	0.06
Sodium, meq/L	137 (135 - 139)	138 (135 - 141)	0.001
Severity of Illness			
SOFA Score	2 (1 – 3)	3 (1 – 5)	<0.001
Medication Exposures			
NSAIDs	27 (8)	19 (6)	0.22
Vasopressors	5 (1)	118 (35)	<0.001
Loop Diuretic	132 (39)	112 (33)	0.10
ACE-Inhibitor / Angiotensin Receptor Blocker	98 (29)	39 (11)	<0.001
Antibiotics	184 (54)	253 (74)	<0.001
Intravenous Contrast	61 (18)	84 (25)	0.03
Randomization Status			
Alert	180 (53)	170 (50)	0.42

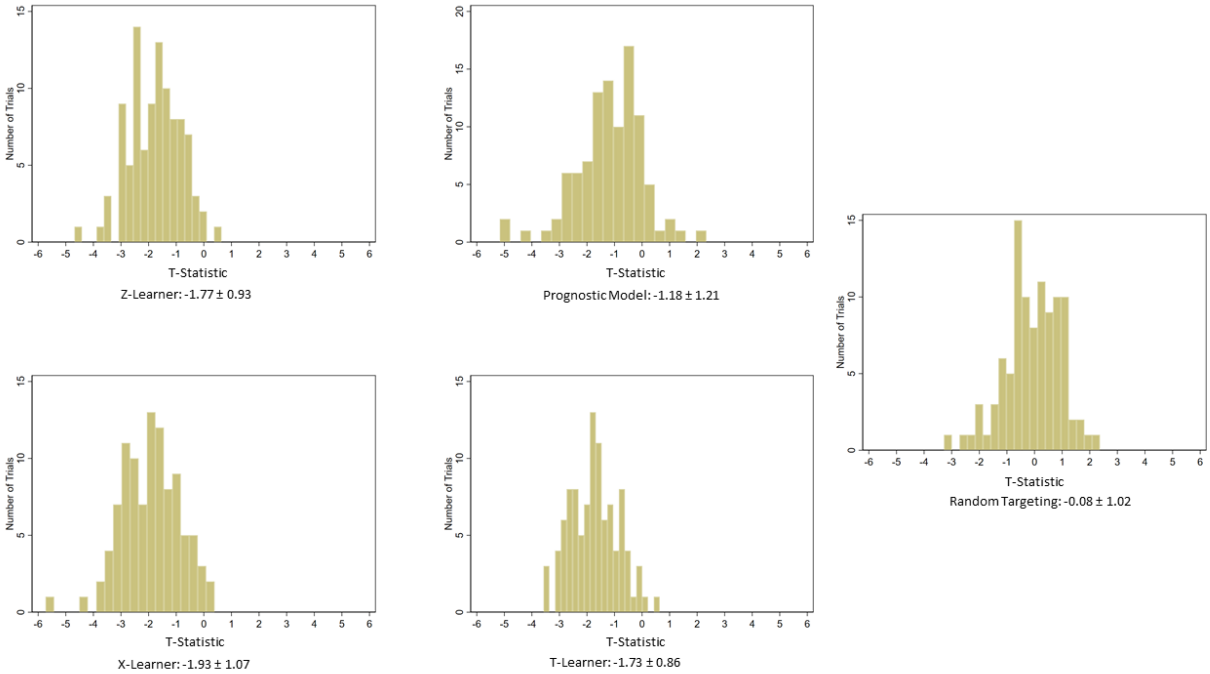
Characteristics at randomization of patients in the test set predicted to be unlikely to benefit from alert vs. likely to benefit from alert based on Prog-method score. Groups are split at the median as the prognostic model does not predict marginal benefit, but ΔCR_3. Note that comorbidity codes and medication exposures are not used in model-building, and are produced here to provide a clearer clinical phenotype for readers.

Supplemental Table 3: Performance of Uplift Algorithms Over 100 Pseudointerventions Applied to a Randomized Trial of Acute Kidney Injury Alerts

Method	Mean (SD) T-Score	Number of trials with T<-1.96
Z-Learner	0.01 (1.09)	0
T-Learner	0.08 (1.03)	2
X-Learner	0.08 (1.08)	1
Prognostic-Learner	-0.03 (0.93)	1

Supplemental Table 3: We evaluated the performance of the uplift models across 100 randomly assigned "pseudointerventions". As these were entirely generated *post hoc*, there can be no subgroup that uniquely benefits from these interventions. Trials with a T-score of <-1.96 would be considered statistically significant evidence of successful targeting. One would expect 2.5 such "successful" examples given 100 pseudointerventions. The mean t-scores of near 0 are also suggestive that the algorithms do not find uplift where none can possibly exist.

Supplemental Figure 1: Distribution of algorithm performance across 100 random training / test splits.



Supplemental Figure 1: Distributions of T-Statistics across 100 random 70% training/ 30% test splits. Values less than 0 imply that the algorithm increases the observed effect size of the intervention in a given test set. All three uplift models almost universally improve the effect of alerting, while the prognostic model does so to a lesser degree. The random model, as expected, would not lead to identification of subgroups who would benefit from an alert.

Supplemental Box 1

Algorithms: Individual Treatment Effect (ITE/Uplift) Estimators

Input is X; *Target* is Y; *Action* is $A \in \{\text{treatment, control}\}$

$ITE(x) := \mathbb{E}[Y|X = x, A = \text{treatment}] - \mathbb{E}[Y|X = x, A = \text{control}] = \delta(x)$

procedure T-LEARNER(X, Y, A)
  Let (X_t, Y_t) be the subgroup of (X, Y) where $A = \text{treatment}$
  Let (X_c, Y_c) be the subgroup of (X, Y) where $A = \text{control}$
  Train a parameterized function f_t to predict Y_t using X_t
  Train a parameterized function f_c to predict Y_c using X_c
  Let $\hat{\delta}(x) = f_t(x) - f_c(x)$ for some x

procedure X-LEARNER(X, Y, A)
  Create a T-learner to obtain: $(X_t, Y_t), (X_c, Y_c), F_t, F_c$
  Calculate residual $R_t = Y_t - f_c(X_t)$
  Calculate negative residual $R_c = -(Y_c - f_t(X_c))$
  Train a parameterized function g_t to predict R_t using X_t
  Train a parameterized function g_c to predict R_c using X_c
  Let $\hat{\delta}(x) = \frac{1}{2}(g_t(x) + g_c(x))$ for some x

procedure Z-LEARNER(X, Y, A)
  Create a new variable Z such that:
    $Z = Y$ if $A = \text{treatment}$
    $Z = -Y$ if $A = \text{control}$
  Train a parameterized function h to predict Z using X
  Let $\hat{\delta}(x) = h(x)$ for some x

procedure PROGNOSTIC-LEARNER(X, Y)
  Train a parameterized function h to predict Y using X
  Let $\hat{\delta}(x) \propto h(x)$ for some x

Supplemental Box 1: Generalized Expression of Uplift Estimators. Methods for calculating uplift (individual treatment effect). X=candidate covariates. Y=Outcome (ΔCr3). A=Treatment assignment. R=Residual. The parameterized functions in this case are linear regressions. Uplift, or individual treatment effect, is an estimate of the marginal benefit of alert as opposed to usual care on ΔCr3.