Supplementary Appendices: Predicting Preventable Hospital Readmissions with Causal Machine Learning

Current version: September, 2020.

Supplementary Information: The Transitions Program Intervention

The Transitions Program intervention consists of a bundle of discrete interventions aimed at improving the transition from inpatient care to home or a to a skilled nursing facility. Together, they comprise a care pathway over the 30 days following discharge, beginning on the morning of the planned discharge day, and which we summarize here, step-by-step. This pathway is also summarized in Table 1 below.

For inpatients awaiting discharge and who have been assigned to be followed by the Transitions Program, on their planned discharge day, a case manager meets with them at the bedside to provide information on the Transitions Program. Next, once the patient has arrived home, a Transitions case manager calls them within 24 to 48 hours to walk them through their discharge instructions, and to identify any gaps in their understanding of them. If necessary, the case manager can also refer the patient to a pharmacist or social worker for focused follow-up. At the same time, the nurse also works to make an appointment with the patient's primary care physician to take place within 3 to 5 days post-discharge.

Following this initial outreach, the Transitions case manager continues to contact the patient weekly by phone, and remains available throughout if the patient requires further assistance. At 30 days post-discharge, the patient is considered to have "graduated" and is no longer followed by the Transitions Program. All steps of this process are initiated through and documented in the EHR, enabling consistent followup for the patients enrolled. A special category of patients are considered at very high risk if their predicted risk is ≥45% or if they lack social support, and receive a more intensified version of the intervention. This version entails follow-up every other day via telephone for the first week post-discharge, followed by ≥2 times a week the second week, and once a week afterward until "graduation" at 30 days.

Risk Level	Initial Assessment	Week 1	Week 2	Week 3	Week 4
High (≥ 45%)	Phone follow-up within 24 to 48 hours *and*	Phone follow-up every other day	2 phone follow-ups; more as needed	Phone follow-up once weekly; more as needed	Phone follow-up once weekly; more as needed
Medium (25 – 45%)	Primary care physician follow-up visit within 2 to 5 days	Once weekly phone follow-up (with more as needed)			
Low (≤ 25%)	Usual care at discretion of discharging physician				

Table 1: The Transitions Program intervention pathway. The initial assessment applies to both the medium and high risk groups. Following it, the pathway diverges in terms of the frequency of phone contact.

Supplementary Information: Technical Appendix

This technical appendix describes in further detail the identification strategy that we undertake, as well as some aspects of causal forests and how to incorporate estimates of "payoffs" into modeling.

TA.1. Identification Strategy

We begin with some notation: for each of a set of units $i = 1, \ldots, n$, we observe the triple (X_i, Y_i, W_i), where $X_i \in \mathbb{R}^p$ is a covariate vector, $Y_i \in \{0, 1\}$ denotes the observed outcome, and $W_i \in \{0, 1\}$ treatment assignment. Following Rubin's potential outcomes framework[1], we assume the existence of potential outcomes, $Y_i(1)$ and $Y_i(0)$, for each unit, and define the conditional average treatment effect (CATE) for an unit i with the covariate vector $X_i = x$ as

$$\tau_i(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x], \tag{1}$$

where $\mathbb{E}$ denotes the expectation operator, $Y_i(1)$ and $Y_i(0)$ the potential outcomes of a 30-day readmission and otherwise, respectively, and x the covariate vector associated with a patient i. Within a leaf L, a causal forest estimates this quantity as

$$\hat{\tau}(x) = \frac{1}{|\{j : W_j = 1 \wedge X_j \in L\}|} \sum_{\{j : W_j = 1 \wedge X_j \in L\}} Y_j \quad - \quad \frac{1}{|\{j : W_j = 0 \wedge X_j \in L\}|} \sum_{\{j : W_j = 0 \wedge X_j \in L\}} Y_j \tag{2}$$

for a $x \in L$. (The $\wedge$ operator represents a logical 'and', and $|A|$ denotes the cardinality of a set A.) Heuristically, the process of fitting a causal forest aims to make these leaves L as small as possible so that the data in each resemble a randomized experiment, while simultaneously maximizing effect heterogeneity.[2]

However, as we observe only one of the two potential outcomes for each unit, $Y_i = Y_i(W_i)$, we cannot estimate $Y_i(1) - Y_i(0)$ directly from these data. Under some assumptions, however, we can reprovision the units in the data that did experience the counterfactual outcome to estimate τ_i, by having those units serve as 'virtual twins' for an unit i. These assumptions entail (1) the existence of these twins; and (2) that, in some sense, these twins look similar in terms of their covariates. These are the *overlap* and *uncounfoundedness* assumptions, respectively. Heuristically, the overlap assumption presumes that these twins could exist, and unconfoundedness posits that these twins are in fact similar in terms of their observed covariates. Together, these assumptions allow us to

have some confidence that the $\tau_i(x)$ in fact do identify causal effects.

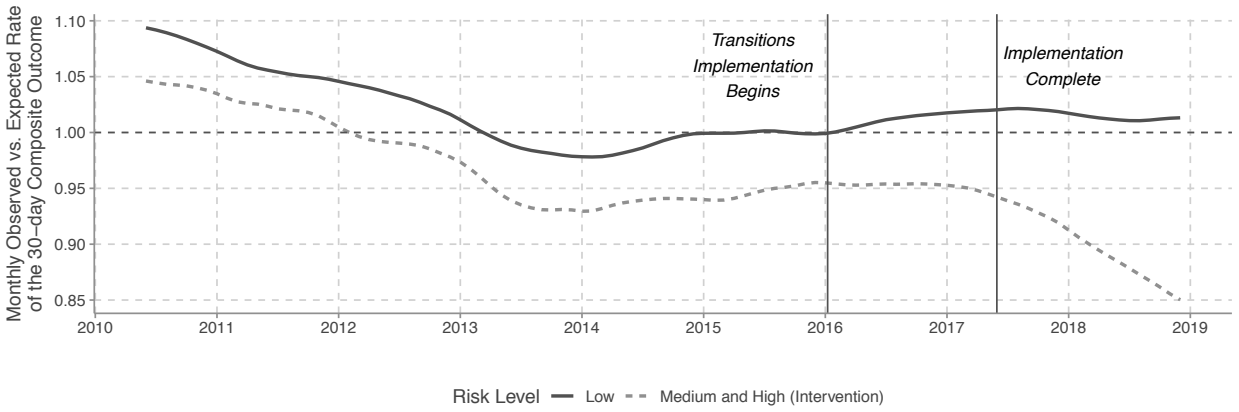

Figure TA1: Trends in the outcome over the study period, 2010-2018, where the data are stratified into comparison and intervention groups. Note the parallel pre-implementation trends, and the substantially similar trend for the comparison group both before and after implementation. Both time series have been deseasonalized using the X-11 method.

We address identification with respect to the predicted risk threshold of 25%, as well as the time period. By time period, recall that the Transitions Program intervention was rolled out to each of the 21 KPNC hospitals in a way that formed two completely disjoint subsets of patients, corresponding to the pre- and post-implementation periods. Also recall that patients were assigned to the intervention if they were discharged during the post-implementation period and their risk score was >25%, while those below that value received usual care. An useful feature of these data is that all patients in the pre-implementation period were assigned 'shadow' risk scores using the same instantiation of the predictive model, even though none of these patients received treatment. (Figure TA1) Hence, the data are split into four disjoint subgroups, indexed by risk category and time period:

$$(\text{pre}, \geq 25), \quad (\text{post}, \geq 25), \quad (\text{pre}, <25), \quad (\text{post}, <25),$$

i.e., the tuple (pre, ≥25) denotes the subgroup consisting of patients discharged during the pre-implementation period with a risk score of ≥25%, and $(.\,, \geq 25)$ denotes all patients with risk ≥25% in the data. Each hospital discharge belongs to one and only one of these subgroups. Importantly, only the patients in the (post, ≥25) subgroup receive the treatment assignment indicator $W_i = 1$. In this respect, our overall approach somewhat resembles a difference-in-differences analysis with the time dimension omitted.

Heuristically, these 'shadow' risk scores allow us to mix data from across periods so that the

(pre, ≥ 25) subgroup can be used as a source of counterfactuals for patients in the (post, ≥ 25) subgroup. Stratifying on the risk score allows us to deconfound the potential outcomes of the patients in these two subgroups. Moreover, these two subgroups together can be used to provide plausible counterfactuals for the patients in the (. , <25) risk subgroup, despite none of those patients having been assigned to the intervention. We describe the identification strategy—which relies on standard ignorability assumptions—in more detail below, beginning with the (. , ≥25) subgroup.

First, for each of the four subgroups, we assume overlap: given some $\epsilon > 0$ and all possible $x \in \mathbb{R}^p$,

$$\epsilon < P(W_i = 1 \mid X_i = x) < 1 - \epsilon. \tag{3}$$

This assumption means that no patient is guaranteed to receive the intervention, nor are they guaranteed to receive the control, based on their covariates x.

For the patients in the (. , ≥25) group, our identification strategy makes use of the balancing properties of risk scores (or prognostic scores)[3] to establish unconfoundedness. Assuming no hidden bias, conditioning on a prognostic score $\Psi(x) = P(Y \mid W = 0, x)$ is sufficient to deconfound the potential outcomes. The risk score used to assign the Transitions intervention is a prognostic score; hence, it is sufficient to deconfound these potential outcomes.

For patients in the (. , <25) subgroup, the picture is slightly more complicated. Among these patients, we cannot assume exchangeability conditional on the risk score $\hat{Y}_i$,

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i \mid \hat{Y}_i, \tag{4}$$

because, again, treatment assignment is contingent on a predicted risk $\hat{Y}_i \geq 0.25$, i.e., $W_i = \mathbf{1}\{\hat{Y}_i \geq 0.25\}$. However, recall that the causal forests are performing estimation in X_i-space, and not in $\hat{Y}_i$-space, and note that we can instead impose the slightly weaker assumption of ignorability conditional on some subset $X_i' \subseteq X_i$, which comprise inputs to the score $\hat{Y}_i$; namely, that

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i \mid X_i', \tag{5}$$

which we can justify *a priori* with the knowledge that no one component predictor predominates in the risk model (see the appendix to[4]); that is, no one covariate strongly determines treatment assignment. We provide empirical evidence to establish the plausibility of this assumption, at

least in low dimensions, in the figure below. Together with the potential outcomes assumption, this *unconfoundedness* assumption is sufficient to obtain consistent estimates of $\tau(x)$[2]. Moreover, since causal forests perform a form of local estimation, our assumptions are independent for the $(., \geq 25)$ and $(., <25)$ subgroups in the sense that if the unconfoundedness assumption fails for either subgroup, but not the other, the estimates for the subgroup in which it does hold should not be affected. Finally, to mitigate bias further, we also estimate propensity scores and an outcome model using regression forests before fitting the causal forest; see the code provided at the end of this Appendix for implementation details.

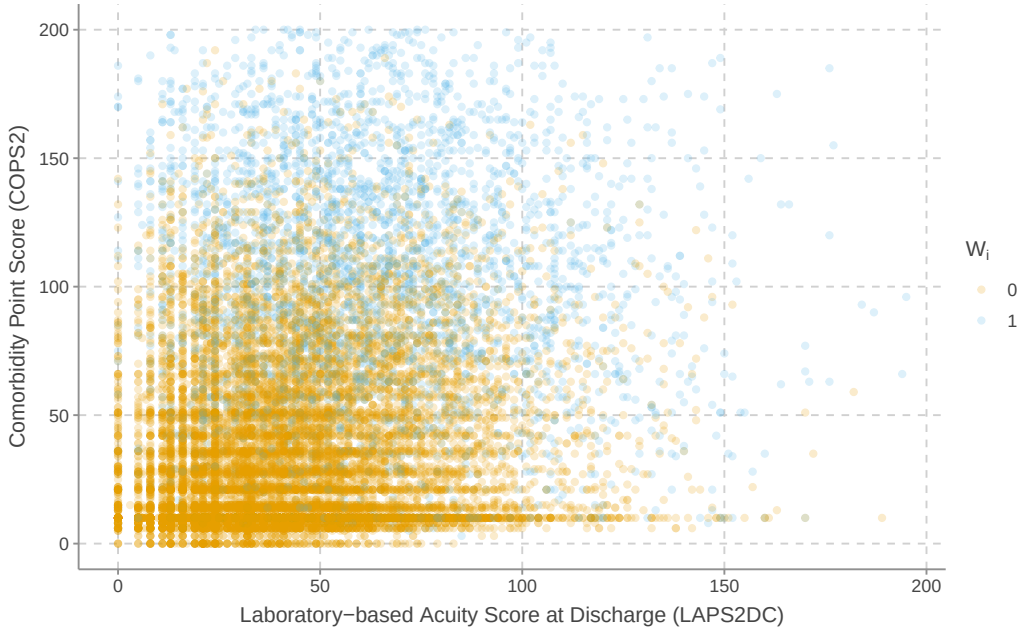

Figure TA2: Assessing the "unconfoundedness" assumption: each point denotes an admission, which are colored according to whether they received the Transitions Program intervention post-discharge (W_i). The x-axis records the value of a laboratory-based acuity score (`LAPS2DC`) and the y-axis the value of a chronic condition score (`COPS2`), both at discharge. The extent of overlap displayed here is relatively good, and implies that overlap may be implausible only among patients at very high or very low risk. This plot is based on a random sample of $n = 20,000$ index admissions taken from the post-implementation period.

In addition, as a formal assessment of treatment effect heterogeneity, we also perform the omnibus test for heterogeneity[5], which seeks to estimate the best linear predictor of the CATE by using the "out-of-bag" predictions from the causal forest, $\hat{\tau}^{-i}$, to fit a linear model.

$$Y_i - \hat{m}^{-i}(X_i) = \alpha\bar{\tau}(W_i - \hat{e}^{-i}(X_i)) + \beta(\hat{\tau}^{-i}(X_i) - \bar{\tau})(W_i - \hat{e}^{-i}(X_i)) + \epsilon, \tag{6}$$

where

$$\bar{\tau} = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}^{-i}(X_i), \tag{7}$$

and $\hat{m}(.)$ and $\hat{e}(.)$ denote the marginal outcome and assignment models estimated by the causal forest, respectively. (The superscript $-i$ denotes that the quantity was computed "out-of-bag", i.e., that the forest was not trained on example i). Fitting this linear model yields two coefficient estimates, α and β; an interpretation of these coefficients is that α captures the average treatment effect, and if $\alpha \approx 1$, then the predictions the forest makes are correct, on average. Likewise, β measures how the estimated CATEs covary with the true CATEs; if $\beta \approx 1$, then these CATE estimates are well-calibrated. Moreover, we can use the p-value for β as an omnibus test for heterogeneity; if the coefficient is statistically significantly greater than zero, then we can reject the null hypothesis of no treatment effect heterogeneity[6].

TA.2. Decoupling causal effects and payoffs

In some cases, the predicted causal effects may not be sufficient to select patients to whom to target such an intervention. The obvious approach starts by treating all patients i with $\hat{\tau}(X_i) < 0$—that is, by treating all patients who are expected to benefit. However, there are two problems with this approach: (1) it is not utility-maximizing, in the sense that it maximizes, for example, the aggregate length of stay (LOS) among the readmissions successfully prevented, and 2), resources may be constrained so that it is infeasible to treat all these patients, making it necessary to prioritize from among those with $\hat{\tau}_i < 0$. One way to do so is to incorporate the costs associated with the potential outcome of a readmission, or the "payoffs" π associated with successfully preventing a readmission[7], which we denote by $\pi_i = \pi(X_i)$.

There are several ways to characterize these payoffs, which ideally can be done mainly in terms of the direct costs required to provide care for a readmitted patient, as well as financial penalties associated with high readmission rates. However, these data are not available to us, so, notionally, we could instead use the LOS of the readmission as a proxy for cost, and assume that higher LOS is associated with higher resource utilization and thus higher costs. It is important to emphasize that these payoffs are associated with the characteristics of the readmission following the index stay, if one does occur—not those of the index stay itself.

One approach to estimating these payoffs is to predict them using historical data, i.e., $\hat{\pi}(X_i) = \mathbb{E}[\pi_i \mid X_i = x]$ in a manner similar to that used to derive the risk scores. However, this is beyond the

scope of this paper, and so we make some simplifying assumptions regarding the payoffs. Namely, we assume that (1) the individual payoffs π_i can be approximated by the mean payoff across all patients, $\pi_i \approx \mathbb{E}[\pi_i]$, and (2) that the payoffs are mean independent of the predicted treatment effects, $\mathbb{E}[\pi_i \mid \tau_i] = \mathbb{E}[\pi_i]$. Figure TA3 establishes the plausibility of Assumption 1 in this setting. These two assumptions make it so that the $\hat{\tau}_i$ become the sole decision criterion for the treatment policies we evaluate in this paper, but we close this section by briefly outlining how to incorporate these payoffs into decision-making if so desired.

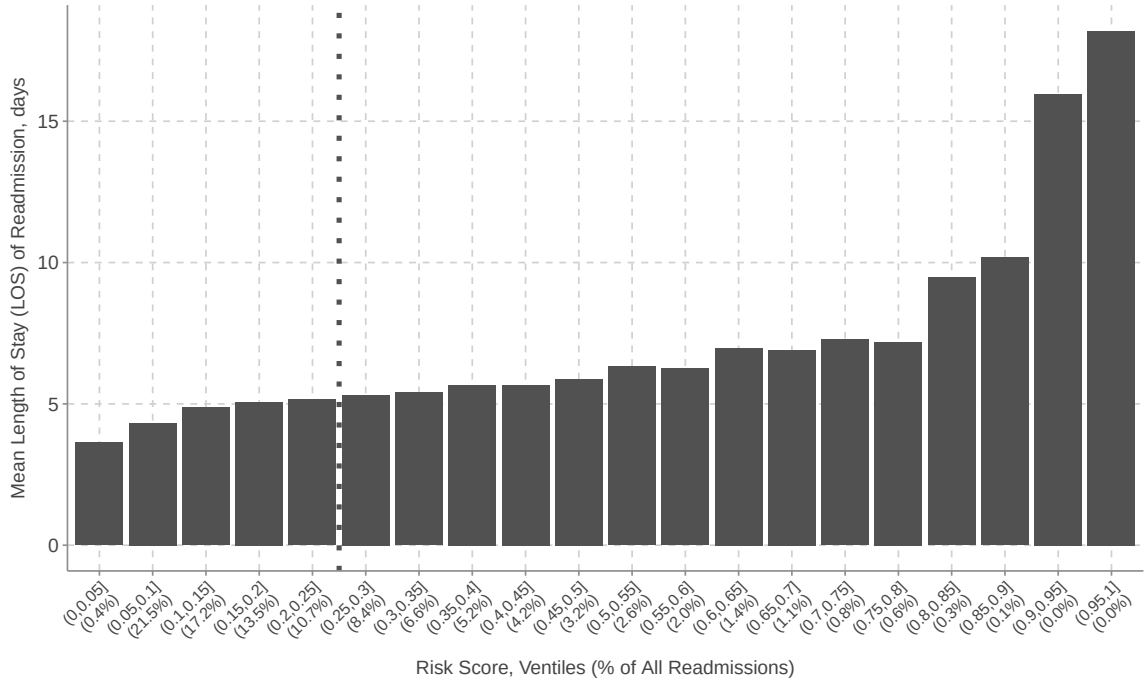

Figure TA3: Average length of stay (LOS) by risk score ventile. The values in parentheses below the name of each ventile denote the proportion of all 30-day readmissions incurred by patients in that ventile; patients with a predicted risk below 25% based on their index stay accounted for 63% of all readmissions. Notably, the average LOS is roughly similar (at 5 days) for patients with predicted risk of 5% to 80%. The vertical dotted line represents the 25% risk threshold used to assign the Transitions Program intervention.

Given both the predicted treatment effects, $\hat{\tau}_i$, and payoffs, $\hat{\pi}_i$, we can compute the individual expected utilities, $\mathbb{E}[u_i] = -\hat{\tau}_i\hat{\pi}_i$ for each patient. We assume that decision-makers are risk-neutral and that the cost to intervene is fixed. Then, given two patients, i and j, and their respective expected utilities, we would prefer to treat i over j if $\mathbb{E}[u_i] > \mathbb{E}[u_j]$. Another interpretation (in a population sense) is that ordering the discharges in terms of their $\hat{\tau}_i$ induces one rank ordering, while ordering them in terms of their $\mathbb{E}[u_i]$ induces another. We can treat the top k% of either ordering, subject to resource constraints, but doing so with the latter will result in greater aggregate (or net) benefit in terms of the chosen units used to characterize the payoffs and thus would be preferred. Under

the assumptions we make above, $\mathbb{E}[u_i] \propto \hat{\tau}_i$ for each patient i. This particular decision-theoretic approach requires absolute, and not relative outcome measures, such as the relative risk reduction.[8]

References

1. Rubin Donald B.. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology.* 1974;66:688–701.

2. Wager Stefan, Athey Susan. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association.* 2018;1459:1–15.

3. Hansen Ben B.. The prognostic analogue of the propensity score. *Biometrika.* 2008;95:481–488.

4. Escobar Gabriel J., Ragins Arona, Scheirer Peter, Liu Vincent, Robles Jay, Kipnis Patricia. Nonelective Rehospitalizations and Postdischarge Mortality. *Medical Care.* 2015;53:916–923.

5. Chernozhukov Victor, Demirer Mert, Duflo Esther, Fernandez-Val Ivan. Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments. 2017.

6. Athey Susan, Wager Stefan. Estimating Treatment Effects with Causal Forests: An Application. *arXiv.* 2019:https://arxiv.org/pdf/1902.07409.pdf.

7. Kleinberg Jon, Ludwig Jens, Mullainathan Sendhil, Obermeyer Ziad. Prediction Policy Problems. *American Economic Review: Papers & Proceedings.* 2015;105:491–495.

8. Sprenger Jan, Stegenga Jacob. Three Arguments for Absolute Outcome Measures. *Philosophy of Science.* 2017;84:840–852.

Supplementary Tables and Figures

	Total	Pre-implementation	Post-implementation	*p*-value	SMD
Hospitalizations, *n*	1,584,902	1,161,452	423,450	—	—
Patients, *n*	753,587	594,053	266,478	—	—
Inpatient (%)	82.8 (69.7-90.6)	84.4 (73.1-90.7)	78.5 (57.7-90.3)	< 0.0001	-0.151
Observation (%)	17.2 (9.4-30.3)	15.6 (9.3-26.9)	21.5 (9.7-42.3)	< 0.0001	0.151
Inpatient stay < 24 hours	5.2 (3.3-6.7)	5.1 (3.8-6.5)	5.6 (1.8-9.1)	< 0.0001	0.041
Transport-in	4.5 (1.4-8.7)	4.5 (1.7-8.8)	4.5 (0.4-8.4)	0.56	-0.001
Age, mean (years)	65.3 (62.2-69.8)	65.1 (61.9-69.6)	65.8 (62.8-70.4)	< 0.0001	0.038
Male gender (%)	47.5 (43.4-53.8)	47.0 (42.4-53.5)	48.9 (45.3-54.9	< 0.0001	0.037
KFHP membership (%)	93.5 (75.3-97.9)	93.9 (80.0-98.0)	92.5 (61.7-97.6)	< 0.0001	-0.052
Met strict membership definition (%)	80.0 (63.4-84.9)	80.6 (67.6-85.5)	78.5 (51.3-83.8)	< 0.0001	-0.053
Met regulatory definition (%)	61.9 (47.2-69.7)	63.9 (50.2-72.2)	56.5 (38.7-66.6)	< 0.0001	-0.152
Admission via ED (%)	70.4 (56.7-82.0)	68.9 (56.0-80.3)	74.4 (58.4-86.6)	< 0.0001	0.121
Charlson score, median (points)	2.0 (2.0-3.0)	2.0 (2.0-3.0)	2.0 (2.0-3.0)	< 0.0001	0.208
Charlson score ≥ 4 (%)	35.2 (29.2-40.7)	33.2 (28.2-39.8)	40.9 (33.0-46.2)	< 0.0001	0.161
COPS2, mean (points)	45.6 (39.1-52.4)	43.5 (38.4-51.5)	51.2 (39.7-55.8)	< 0.0001	0.159
COPS2 ≥ 65 (%)	26.9 (21.5-32.0)	25.3 (21.0-31.6)	31.1 (22.5-35.4)	< 0.0001	0.129
Admission LAPS2, mean (points)	58.6 (48.0-67.6)	57.6 (47.4-65.8)	61.3 (50.2-72.8)	< 0.0001	0.092
Discharge LAPS2, mean (points)	46.7 (42.5-50.8)	46.3 (42.5-50.8)	47.6 (42.3-52.9)	< 0.0001	0.039
LAPS2 ≥ 110 (%)	12.0 (7.8-16.0)	11.6 (7.5-15.2)	12.9 (8.3-18.4)	< 0.0001	0.039
Full code at discharge (%)	84.4 (77.3-90.5)	84.5 (77.7-90.5)	83.9 (75.9-90.5)	< 0.0001	-0.016
Length of stay, days (mean)	4.8 (3.9-5.4)	4.9 (3.9-5.4)	4.7 (3.9-5.6)	< 0.0001	-0.034
Discharge disposition (%)					0.082
To home	72.7 (61.0-86.2)	73.3 (63.9-85.9)	71.0 (52.1-86.9)	< 0.0001	
Home Health	16.1 (6.9-23.3)	15.2 (6.9-22.6)	18.5 (7.0-34.5)	< 0.0001	
Regular SNF	9.9 (5.9-14.3)	10.0 (6.0-15.2)	9.5 (5.6-12.4)	< 0.0001	
Custodial SNF	1.3 (0.7-2.5)	1.5 (0.8-2.7)	0.9 (0.4-1.8)	< 0.0001	
Hospice referral (%)	2.6 (1.7-4.4)	2.6 (1.7-4.6)	2.7 (1.5-4.0)	< 0.0001	0.007
Outcomes					
Inpatient mortality (%)	2.8 (2.1-3.3)	2.8 (2.1-3.3)	2.8 (1.8-3.3)	0.17	-0.003
30-day mortality (%)	6.0 (4.0-7.3)	6.1 (4.1-7.6)	5.9 (3.9-6.8)	< 0.0001	-0.006
Any readmission (%)	14.5 (12.7-17.2)	14.3 (12.3-17.3)	15.1 (13.3-17.0)	< 0.0001	0.021
Any non-elective readmission (%)	12.4 (10.4-15.4)	12.2 (10.2-15.5)	13.1 (10.8-15.4)	< 0.0001	0.029
Non-elective inpatient readmission (%)	10.5 (8.2-12.6)	10.4 (8.1-12.8)	10.8 (8.6-12.9)	< 0.0001	0.012
Non-elective observation readmission (%)	2.4 (1.4-3.7)	2.2 (1.2-3.4)	3.0 (1.9-5.6)	< 0.0001	0.049
30-day post-discharge mortality (%)	4.0 (2.6-5.2)	4.1 (2.7-5.4)	3.9 (2.3-4.9)	< 0.0001	-0.007
Composite outcome (%)	15.2 (12.9-18.8)	15.0 (12.9-19.1)	15.8 (13.3-18.0)	< 0.0001	0.023

Table S1: Characteristics of the cohort, including both index and non-index stays. Notably, comparing pre- to post-implementation, hospitalized patients were older, and tended to have higher comorbidity burden (higher COPS2) as well as a higher acuity of illness at admission (higher LAPS2). The use of observation stays also increased. These differences reflect a broader trend towards the pool of potential inpatient admissions becoming more and more ill over the decade from 2010, in large part due to the effectiveness of outpatient preventative care processes at KPNC, as well as of programs providing care outside of the hospital setting as an alternative to admission. Otherwise, care patterns did not substantially change, as evidenced by, e.g., transports-in, Kaiser Foundation Health Plan (KFHP) membership status, and discharge disposition mix, all of which had standardized mean differences (SMDs) < 0.1. In large cohorts such as this one, SMDs can be a better guide to detecting covariate imbalances or differences between groups, owing to the effects of large sample sizes. Finally, as a consequence of increased comorbidity burden and admission acuity, and despite the implementation of the Transitions Program, rates of readmission and of the composite outcome increased from pre- to post-implementation. Abbreviations: SMD, standardized mean difference; KFHP, Kaiser Foundation Health Plan; LAPS2, Laboratory-based Acute Physiology Score, version 2; COPS2, COmorbidity Point Score, version 2; SNF, skilled nursing facility.

Supergroup name (HCUPSGDC)	Clinical Classification Software (CCS) category code(s)
Acute CVD	109
AMI	100
CAP	122
Cardiac arrest	107
CHF	108
Coma; stupor; and brain damage	85
Endocrine & related conditions	48-51, 53, 54, 56, 58, 200, 202, 210, 211
Fluid and electrolyte disorders	55
GI bleed	153
Hematologic conditions	59-64
Highly malignant cancer	17, 19, 27, 33, 35, 38-43
Hip fracture	226
Ill-defined signs and symptoms	250-253
Less severe cancer	11-16, 18, 20-26, 28-32, 34, 36, 37, 44-47, 207
Liver and pancreatic disorders	151, 152
Miscellaneous GI conditions	137-140, 155, 214
Miscellaneous neurological conditions	79-84, 93-95, 110-113, 216, 245, 653
Miscellaneous surgical conditions	86-89, 91, 118-121, 136, 142, 143, 167, 203, 204, 206, 208, 209, 212, 237, 238, 254, 257
Other cardiac conditions	96-99, 103-105, 114, 116, 117, 213, 217
Other infectious conditions	1, 3-9, 76-78, 90, 92, 123-126, 134, 135, 148, 197-199, 201, 246-248
Renal failure (all)	156, 157, 158
Residual codes	259
Sepsis	2
Trauma	205, 225, 227-236, 239, 240, 244
UTI	159

Table S2: List of Clinical Classification Software (CCS)-defined supergroups and their CCS codes used in this study. These supergroups represent levels of the covariate HCUPSGDC. More details on the CCS codes themselves, as well as mappings to their component ICD codes, can be found at www.ahrq.gov/data/hcup.

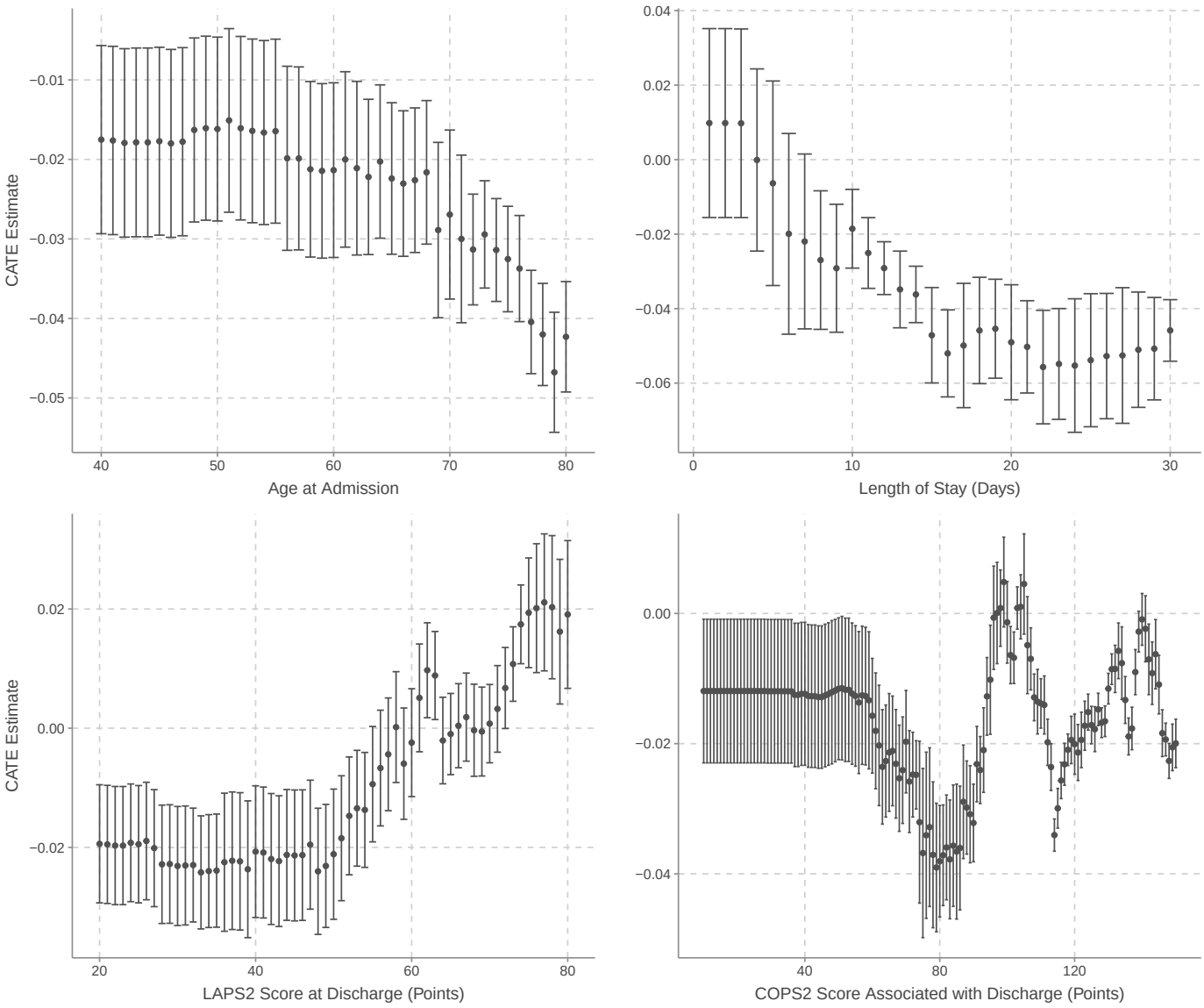

Figure S1: Additional dimensions of heterogeneity as visualized by the estimated CATE function. These figures show treatment effect heterogeneity among some important covariates, including patient age at admission, length of stay (in days) of their index stay, LAPS2 score at discharge, and the COPS2 score associated with the index stay. These resemble Figure 3 in the main manuscript, and take the same 'pseudo-patient' approach, but use only one dimension (i.e., one covariate) as opposed to two. Where applicable, the supergroup was set to heart failure, gender to female, age on admission to 60, length of stay to 5 days, LAPS2 on admission and at discharge to 60 and 45, respectively, and COPS2 score to 50. Error bars represent confidence intervals.

Listing 1: Code used to produce the analyses.

```
library(grf)
library(ggplot2)
library(zoo)
library(ggridges)
library(dplyr)
library(tidyr)

# Assumes matrix X, and vectors W and Y
# X: covariate matrix; W and Y indicate treatment assignment and outcome (30-day readmission), respectively.

# Outcome model, marginalizing over W (m(x))
Y.forest <- regression_forest(X, Y, clusters = facility.ids)
Y.hat <- predict(Y.forest, X)$predictions

# Propensity model
W.forest <- regression_forest(X, W, clusters = facility.id)
W.hat <- predict(W.forest, X)$predictions

cf.main <- causal_forest(X,
                         Y,
                         W,
                         Y.hat = Y.hat,
                         W.hat = W.hat,
                         clusters = facility.id,
                         num.trees = 8000,
                         min.node.size = 10,
                         tune.parameters = TRUE)

# Omnibus test for heterogeneity
test_calibration(cf.main)

# ATE (via target.sample = 'all') (think of as average CATE)
average_treatment_effect(cf.raw, target.sample = 'all')

# Get out-of-bag predictions for all patients.
oob.preds <- predict(cf.main, estimate.variance = TRUE)

# `dataset` refers to main dataset; merge predicted CATEs back in
dataset <- cbind(dataset, oob.preds)

# Table 2: Estimating impact on '18 data
# Assume datasets `dataset.upto2017` and `dataset.2018` -> X.17, W.17, Y.17, and X.18, W.18, Y.18, respectively
# All *.17s represent data up to and including 2017, while 2018 includes data only from 2018.

Y.forest.17 <- regression_forest(X.17, Y.17, clusters = facility.id)
Y.hat.17 <- predict(Y.forest.17, X.17)$predictions

# This functions like a propensity score
W.forest.17 <- regression_forest(X.17, W.17, clusters = facility.id)
W.hat <- predict(W.forest.17, X.17)$predictions

cf.17 <- causal_forest(X.17,
                       Y.17,
                       W.17,
                       Y.hat = Y.hat.17,
                       W.hat = W.hat.17,
                       clusters = facility.id,
                       num.trees = 8000,
                       min.node.size = 10,
                       tune.parameters = TRUE)
```

```
oob.preds.18 <- predict(cf.17, X.18, estimate.variance = TRUE)

dataset.2018 <- cbind(dataset, oob.preds.18)

# Comparing versus baseline: take all >25% risk - note that CATEs are asymptotically normal
baseline.data <- subset(dataset.2018, risk.score >= 0.25)
n.readmits.prevented.baseline <- -sum(baseline.data$predictions) # note minus sign
lower95.readmits.prevented.baseline <- n.readmits.prevented - 1.96 * sqrt(sum(baseline.data$variance.estimates))/nrow(baseline.data)
upper95.readmits.prevented.baseline <- n.readmits.prevented + 1.96 * sqrt(sum(baseline.data$variance.estimates))/nrow(baseline.data)
nnt.readmits.prevented.baseline <- -1/mean(baseline.data$predictions) # note minus sign

# N.B. `risk.score.cut`s are produced with cut(..., breaks=seq(0, 1, by=0.05)) on the risk score.
table2.data <- dataset.2018 %>%
  group_by(risk.score.cut) %>%
  filter(pred.cate < quantile(pred.cate, 0.5)) %>% # n.b.: can repeat for different values of threshold
  ungroup()

# statistics for each row of Table 2
n.readmits.prevented <- -sum(table2.data$predictions) # note minus sign
lower95.readmits.prevented <- n.readmits.prevented - 1.96 * sqrt(sum(table2.data$variance.estimates))/nrow(table2.data)
upper95.readmits.prevented <- n.readmits.prevented + 1.96 * sqrt(sum(table2.data$variance.estimates))/nrow(table2.data)
nnt.readmits.prevented <- -1/mean(table2.data$predictions) # note minus sign

# Figure 1: HTE by ventile.
ggplot(dataset, aes(x=pred.cate, y=risk.score.cuts)) +
  geom_density_ridges(panel_scaling=FALSE) +
  coord_flip() +
  xlab('Out-of-Bag CATE Estimate') +
  ylab('Risk Score, Ventiles')

# Figure 2: HTE stratified by HCUP supergroups
ggplot(dataset, aes(x=pred.cate, y=risk.score.cuts)) +
  geom_density_ridges(panel_scaling=FALSE) +
  coord_flip() +
  xlab('Out-of-Bag CATE Estimate') +
  ylab('Risk Score, ventiles') +
  facet_wrap(HCUPSG_DC ~ .)

# Figure 3: Visualizing CATE function surface for a grid of pseudopatients.

# Pick the first row to retain column layout
X.test <- X[1,]

# Set age at admission to 50 and prior hospitalizations to 0
X.test$AGE_AT_ADMIT <- 50
X.test$hosp_prior7_ct <- 0
X.test$hosp_prior8to30_ct <- 0

# Set to mean LAPS2 on admission for KPNC patients
X.test$LAPS2 <- 55

# Set LOS to mean LOS at D/C
X.test$LOS_30 <- 5

# Set supergroup to CHF (internally 1080)
X.test$HCUPSG_DC_1020 <- 0
X.test$HCUPSG_DC_1080 <- 1

# Create grid of COPS2 and LAPS2 values
COPS2.grid <- seq(10, 150, by=5)
```

```
LAPS2DC.grid <- seq(24, 84, by=6)
l.by.c <- crossing(LAPS2DC.grid, COPS2.grid)

X.test <- X.test %>%
            select(-LAPS2DC, -COPS2)

X.test <- cbind(X.test, l.by.c)
X.test$LAPS2DC <- X.test$LAPS2DC.grid
X.test$COPS2 <- X.test$COPS2.grid

X.test <- X.test %>%
            select(-LAPS2DC.grid, -COPS2.grid)

X.test <- X.test %>%
            select(AGE_AT_ADMIT,
                   MALE,
                   DCO_4,
                   hosp_prior7_ct,
                   hosp_prior8to30_ct,
                   LOS_30,
                   MEDICARE,
                   DISCHDISP,
                   LAPS2,
                   LAPS2DC,
                   COPS2,
                   starts_with('HCUPSG_DC'))

X.test.50 <- X.test
X.test.80 <- X.test
X.test.80$AGE_AT_ADMIT <- 80

pred.50 <- predict(cf.main, X.test.50, estimate.variance = TRUE)
X.pred.res.50 <- cbind(X.test.50, pred.50)

pred.80 <- predict(cf.main, X.test.80, estimate.variance = TRUE)
X.pred.res.80 <- cbind(X.test.80, pred.80)

X.comb <- rbind(X.pred.res.50, X.pred.res.80)

# Make titles for subplots w/ facet_wrap()
X.comb$AGE_disp <- ifelse(X.comb$AGE_AT_ADMIT == 80, 'Age = 80', 'Age = 50')

# Set legend min/max and zero
breaks <- c(min(X.comb$predictions), 0, max(X.comb$predictions))
breaks <- round(breaks, 3)

ggplot(X.comb, aes(LAPS2DC, COPS2)) +
  geom_raster(aes(fill=predictions), interpolate=FALSE) +
  scale_fill_gradient2(breaks=breaks) +
  facet_wrap(AGE_disp ~ ., ncol = 2, nrow=1) +
  labs(fill = "Predicted CATE") +
  xlab('Laboratory-based Acuity Score at Discharge (LAPS2DC)') +
  ylab('Comorbidity Point Score (COPS2)')
```