# How Do Machine Learning Algorithms Perform in Predicting Hospital Choices? Evidence from Changing Environments*

**Devesh Raval**
Federal Trade Commission
draval@ftc.gov

**Ted Rosenbaum**
Federal Trade Commission
trosenbaum@ftc.gov

**Nathan E. Wilson**
Federal Trade Commission
nwilson@ftc.gov

April 14, 2021

**Abstract**

Researchers have found that machine learning methods are typically better at prediction than econometric models when the choice environment is stable. We study hospital demand models, and evaluate the relative performance of machine learning algorithms when the choice environment changes substantially due to natural disasters that closed previously available hospitals. While machine learning algorithms outperform traditional econometric models in prediction, the gain they provide shrinks when patients' choice sets are more profoundly affected. We show that traditional econometric methods provide important additional information when there are major changes in the choice environment.

JEL Codes: C18, I11, L1, L41
Keywords: machine learning, hospitals, natural experiment, patient choice, prediction

*The views expressed in this article are those of the authors. They do not necessarily represent those of the Federal Trade Commission or any of its Commissioners. We are grateful to Jonathan Byars, Gregory Dowd, Aaron Keller, Laura Kmitch, and Peter Nguon for their excellent research assistance. We also thank Chris Garmon, Marty Gaynor, Dan Hosken, Nate Miller, Harry Paarsch, Dave Schmidt, and participants at Auburn, the 2017 IIOC, 2018 ASHEcon Meetings, 2018 Conference on Health IT and Analytics, and 2018 Bank of Canada Data Science Conference for their comments. We also thank anonymous referees from the ACM EC'19 conference.

# 1 Introduction

The proliferation of rich consumer-level datasets has led to the rise of the "algorithmic modeling culture" (Breiman, 2001b), wherein analysts treat the statistical model as a "black box" and predict choices using algorithms trained on existing datasets. Agrawal et al. (2018) predict that reductions in the cost of prediction due to the increasing adoption of machine learning models will revolutionize how businesses address the problems they face. The excitement about new AI methods is, in part, due to the belief that they provide better predictive accuracy than traditional econometric methods.

However, evaluating health policy questions often involves contemplating a substantial shift in the choice environment. For example, a health insurance reform may change the set of insurance products that consumers can buy and provider entry and exit alters the set of products available to patients. For such questions, it is less obvious whether purely data-driven machine learning methods can usefully be applied, compared to models that incorporate domain knowledge through economic assumptions. As Athey (2017) remarks:

> [M]uch less attention has been paid to the limitations of pure prediction methods. When SML [supervised machine learning] applications are used "off the shelf" without understanding the underlying assumptions or ensuring that conditions like stability [of the environment] are met, then the validity and usefulness of the conclusions can be compromised.

In this paper, we use a major change in the choice environment to compare the performance of an econometric hospital demand model to machine learning models. Hospital demand models are widely used for evaluating counterfactual changes in choice sets in hospital mergers (Capps et al., 2003; Farrell et al., 2011; Gowrisankaran et al., 2015; Gaynor et al., 2015), insurance mergers (Ho and Lee, 2019), and narrow insurance networks (Ghili, 2016; Ho and Lee, 2019). To gauge the different models' performance, we use a set of natural disasters that closed one or more hospitals but left the majority of the surrounding area relatively undisturbed. These "shocks" exogenously altered consumers' choice sets, creating a benchmark – patients' actual choices in the post-disaster period – against which to

assess the performance of different predictive models calibrated on pre-disaster data. Our main prediction criterion is the fraction of actual choices that we correctly predict using the highest estimated probability as the predicted choice. By comparing the different models' predictions to actual post-disaster choices, we are able to gauge predictive performance when the choice environment has changed.

Relative to a benchmark econometric choice model akin to those used in recent academic work (Ho, 2006; Gowrisankaran et al., 2015), we consider the performance of ML models that are heavily used by practitioners (Athey and Imbens, 2019) and that are currently implemented in standard software packages. In particular, we compare examples of two classes of machine learning algorithms – grouping and regularization. Grouping models partition the space of patients into types and estimate choice probabilities separately for each type. In this category, we evaluate an "exogenous" grouping model based upon Raval et al. (2017), an individual decision tree model, and two methods, random forests and gradient boosted trees, that are known to improve prediction performance by aggregating over multiple trees. Regularization models involve building a "punishment" term into the objective function that leads to the exclusion of variables that add relatively little new information. In this category, we focus attention on a regularized version of a multinomial logit model that selects the variables most relevant for predicting hospital choices.

We find that the gradient boosting and random forest methods estimated on pre-disaster data generally outperform all other approaches at predicting patient choice after a disaster has closed a hospital. Averaging across all six experiments, the random forest, gradient boosting, and regularization models all correctly predict 46% of choices. By contrast, the benchmark econometric model correctly predicts 40% of choices, while assigning all choices to the highest share hospital in the destroyed hospital's service area correctly predicts 29% of choices. Either the random forest or gradient boosting model is the best predicting model for all of the experiments, and they are the best two models for four of the six experiments.

While it would be hard to objectively distinguish between the performance of the random forest and gradient boosting models, we do find a large difference between these best predicting models in terms of computational time. For our largest dataset, the random forest takes minutes to run while the gradient boosting model takes several hours. The next best

model, the regularized logit, takes almost a week for the same dataset.

We further show that the better performance of machine learning models is not driven by post-disaster changes in patient composition or preferences. Across disasters, the number of admissions falls by 6 to 14% after the hospital is destroyed, which indicates that some patients may either delay or decline treatment. However, we find that the different models perform qualitatively similarly after removing areas that faced more destruction from the disaster and when restricting our sample to cardiac or pregnancy patients who likely have a much more limited ability to delay treatment. We also find similar results for patients with different levels of disease acuity and who have different payers, which further suggests that our results are robust across different patient populations.

In most situations, an analyst will not have an experiment to use to evaluate model performance. Instead, they will only be able to gauge accuracy by "holding out" a portion of their data, and testing how well different models estimated on the remainder of the data do in predicting outcomes in the hold out sample. We find that predictive accuracy in a validation sample formed by holding out 20% of the training data provides a good guide for which models do best at predicting choices after the disasters.

While we consistently find that the machine learning methods perform best at prediction on average, their relative performance deteriorates for patients who were more likely to have had a major change in their choice set. We show this by considering patients who were especially likely to have gone to the destroyed hospital, either because they previously went there or because we predict a high probability of them going there. On average, the relative performance of the machine learning methods over the benchmark logit falls for patients who were more likely to have used the destroyed hospital. For the experiment where the destroyed hospital had a 50% share of the market pre-disaster, all of the machine learning models perform *worse* than our benchmark conditional logit for patients predicted to have a 50% or greater probability of going to the destroyed hospital.

The machine learning models could perform relatively worse with a larger change in the choice set for two main reasons. First, a less local, simpler model with less variable estimates may be required. We test this explanation by varying the minimum node size of the random forest; with a larger minimum node size, the random forest model is less local. However, we

find that random forest models that are less local, and use more patients to estimate each set of probabilities, perform (weakly) worse with a larger change in the choice set.

Second, there may be a greater need to complement the data with the researcher's prior domain knowledge on model specification. In our setting, domain knowledge is reflected by specifying the logit model's parametric form. The econometric model we estimate imposes the parametric restriction that any horizontal, spatial differentiation enters through consumers' travel time to hospitals, as in the canonical model of Hotelling (1929).[1] We quantify the role this domain knowledge may play through an optimal model combination exercise that allocates weights to different models. We find that the weight on the conditional logit model rises as we move from using out of sample validation data for which patients see no change in choice set, to the test sample of post-disaster patients, to subsets of test sample patients with a high probability of visiting the destroyed hospital.

Overall, our work connects to the literature on hospital competition and how to infer providers' substitutability (Gaynor et al., 2015, 2013). Within this literature, it is most closely related to Raval et al. (2020), which shows that econometric demand models may often underpredict the aggregate levels of patient substitution (i.e., "diversion ratios") to hospitals with the the highest observed substitution following natural disasters. While that paper studies predictions about diversion ratios among a set of econometric models, this paper studies predictions about individual choices and compares machine learning models to econometric models.

Outside of the health literature, our work contributes to the emerging literature in economics and quantitative social science on the application of machine learning techniques. Within this literature, our work is closest to Bajari et al. (2015a,b) and Rose (2016), which also focus on evaluating the relative performance of machine learning models given a stable choice environment. These papers consider the out-of-sample performance of machine learning models relative to econometric models of consumer goods demand and health care expenditures, respectively. In contrast, our work studies out-of-sample performance when there are plausibly exogenous changes in the choice environment.

[1] As Raval and Rosenbaum (forthcoming) discuss, spatially heterogeneous preferences for hospitals can come both from differences in consumer travel costs and from other preferences correlated with travel time.

By illustrating how standard machine learning approaches perform in making predictions following a change in the choice environment, our paper also contributes to the growing body of work focused on the proper application of machine learning methods to make causal predictions (Athey, 2017). In contrast to the literature on estimating treatment effects using machine learning (e.g., Belloni et al., 2014; Wager and Athey, 2018), we do not use the variation from the policy change in our estimation. Rather, we estimate a model without using that variation and assess the quality of the models predictions following a change in the environment. In estimating the model without using variation on the change in the environment, we more closely mimic the problem that is frequently faced by policy makers and businesses when making decisions. They need to make decisions where they only have access to information prior to the change occurring.

The paper proceeds as follows. Section 2 discusses our data and experimental settings. Then, in Section 3, we describe the different models we test. Section 4 examines the computational time required for the machine learning algorithms, Section 5 presents the results on model performance, and Section 6 examines how model performance deteriorates for patients experiencing a greater change in environment. Finally, we discuss lessons that practitioners may take from our work and conclude in Section 7.

# 2 Natural Experiments

## 2.1 Disasters

We exploit the unexpected closures of six hospitals in four different regions following three different types of natural disaster. Table 1 below lists the locations of the disasters, when they took place, the nature of the event, and the hospital(s) affected. Our sample includes disasters affecting urban markets (New York City and Los Angeles) as well as rural markets, and elite academic medical centers (NYU Langone) as well as community health centers. Because of this considerable heterogeneity in the "treated" groups, we have broad confidence in the external validity of our results.

Table 1: Natural Disasters

| Location | Month/Year | Severe Weather | Hospital(s) Closed |
|---|---|---|---|
| Northridge, CA | Jan-94 | Earthquake | St. John's Hospital |
| Americus, GA | Mar-07 | Tornado | Sumter Regional Hospital |
| New York, NY | Oct-12 | Superstorm Sandy | NYU Langone |
| | | | Bellevue Hospital Center |
| | | | Coney Island Hospital |
| Moore, OK | May-13 | Tornado | Moore Medical Center |

## 2.2 Service Areas and Choice Sets

Like much of the prior literature, we estimate demand for hospitals for those patients seeking inpatient care using the discharge data collected by state departments of health.[2] Such patient-hospital data have frequently been used by researchers (Capps et al., 2003; Ciliberto and Dranove, 2006; Garmon, 2017). They include many characteristics describing the patient receiving care such as age, sex, zip code of residence, and diagnosis.[3]

To assess the performance of different predictive methods when consumers' choice sets change, we first identify the patient population exposed to the loss of a choice. We do this by constructing the 90% service area for each destroyed hospital using the discharge data. The service area is defined as the smallest set of zip codes that accounts for at least 90% of the hospital's admissions. Because this set may include many zip codes where the hospital is competitively insignificant, we exclude any zip code where the hospital's share in the pre-disaster period is below 4%. We assume that any individual that lived in this service area and required care at a general acute care hospital would have considered the destroyed hospital as a possible choice. We define the set of relevant substitute hospitals as those that have a share of more than 1% of the patients in the 90% service area, as defined above, in a given month (quarter for the smaller Sumter and Moore datasets). We combine hospitals not meeting this threshold into an "outside option."

We estimate the models on data from the period before the disaster, and test them on

[2]For the most part, we rely on data provided directly from the relevant state agencies. For New York, we use both data provided by the state agency and discharge data separately obtained from HCUP. The HCUP data allow us to observe whether a patient had visited the destroyed hospital in the recent past, which we exploit in Section 6 to examine previous patients of the destroyed hospital.

[3]Precise details on the construction of our estimation samples are provided in Appendix B.

admissions taking place after the disaster. We refer to the data from the time period before the disaster as the "training data" and after the disaster as the "test data." We exclude the period immediately surrounding the disaster to avoid including injuries from the disaster and to ensure that the choice environment resembles the pre-period as much as possible.[4]

Table 2 displays characteristics of each destroyed hospital's market environment, including the number of admissions before and after the disaster, the share of the service area that went to the destroyed hospital before the disaster, the number of zipcodes in the service area, and the number of rival hospitals. We also indicate the average acuity of patients choosing the destroyed hospital during the pre-disaster period, measured by average MS-DRG weight.[5]

Table 2 indicates that the service area for Sumter Regional experienced a massive change from the disaster; the share of the destroyed hospital in the service area was over 50 percent. For the other disasters, the disruption was smaller though still significant as the share of the destroyed hospital in the service area ranges from 9 to 18 percent. Thus, the destroyed hospitals consistently have a large enough share in each service area that patients' choice environments are likely to have changed substantially. Table 2 also shows that we have a substantial number of patient admissions before and after each disaster with which to estimate and test the different models. The number of admissions in the training and test datasets ranges from the thousands for Moore and Sumter to tens of thousands for the New York hospitals and St. John's.[6]

[4]Except for St. Johns, the omitted period is just the the month of the disaster. We describe the specific periods dropped for each disaster in Appendix B.

[5]DRG weights are designed to measure the complexity of patients' treatments, so reporting average weights are a way of measuring variation in treatment complexity between hospitals and regions.

[6]The New York service areas do overlap. The service area for NYU is much larger than Bellevue, so most of the zip codes for Bellevue are also in the service area for NYU, but the reverse is not true. NYU has a 3.9 percent share in the Coney service area and 9.5 percent share in the Bellevue service area, and Bellevue has a 5.7 percent share in the NYU service area.

Table 2: Descriptive Statistics of Affected Hospital Service Areas

| | Training Data Admissions | Test Data Admissions | Zip Codes | Choice Set Size | Destroyed Share | Destroyed Acuity |
|---|---|---|---|---|---|---|
| Sumter | 6,940 | 5,092 | 11 | 15 | 50% | 1.02 |
| NYU | 79,950 | 16,696 | 38 | 19 | 9% | 1.41 |
| Coney | 46,588 | 9,666 | 8 | 17 | 18% | 1.16 |
| Bellevue | 46,260 | 9,152 | 19 | 20 | 11% | 1.19 |
| Moore | 9,763 | 3,920 | 5 | 12 | 11% | 0.91 |
| St. Johns | 97,030 | 18,130 | 29 | 21 | 17% | 1.30 |

**Note:** The first column indicates the number of admissions in the (pre-period) training data, the second column the number of admissions in the (post-period) test data, the third column the number of zip codes in the service area, the fourth column the number of choices (including the outside option), the fifth column the share of admissions in the pre-period from the 90% service area that went to the destroyed hospital, and the sixth column the average DRG weight of admissions to the destroyed hospital in the training data.

# 3 Models

Economists have typically modeled hospital demand using a discrete choice framework that conditions on a patient having chosen to receive inpatient hospital care.[7] The econometrician then presumes that patient $i$'s utility from receiving care from each relevant hospital $h$ is a linearly separable combination of a deterministic component based on observable elements $\delta_{ih}$ and an idiosyncratic shock $\epsilon_{ih}$:

$$u_{ih} = \delta_{ih} + \epsilon_{ih}. \tag{1}$$

Since the full set of hospitals may be large, as discussed earlier, we normalize some hospitals to the outside option $h = 0$, with $\delta_{i0} = 0$ for all patients $i$. In addition, the hospital choice literature has generally assumed that $\epsilon_{ih}$ is distributed Type I extreme value (e.g., Capps et al., 2003; Gowrisankaran et al., 2015; Ho and Lee, 2019).

Given the linear separability and distributional assumptions, the fundamental question for the econometrician is how to specify $\delta_{ih}$. All of the estimation approaches we explore can be described as different ways of parameterizing $\delta_{ih}$.

[7]The assumption is that deferring inpatient care is difficult. We address the impact of violations of this assumption on our results in Section 5.3.

## 3.1 Models of Patient Choice from the Econometric Literature

In the empirical economics literature on hospital choice, economists ex-ante specify models for deterministic utility $\delta_{ih}$. Although the models used in the literature vary in what explanatory terms they include, they make two basic assumptions on consumer choice. First, patients care about how costly, in terms of travel time, it is for them to receive care at different hospitals (Hotelling, 1929), which provides a source of horizontal differentiation in hospital preferences. Second, hospitals are observably vertically differentiated in their appeal to consumers on quality. Different models may allow for both preferences over travel time, and hospital quality, to be differentially attractive to different patient types.

These models represent variants of the following general form:

$$\delta_{ih} = \sum_{k} \gamma_{kh} z_{ik} \alpha_h + f(d_{ih}, z_{ik}, \alpha h), \tag{2}$$

where $i$ indexes patient, $h$ indexes hospital, and $k$ indexes patient characteristics. Then, $z_{ik}$ are patient characteristics (e.g., age, condition, location, etc.), $\alpha_h$ are hospital indicators (alternative specific constants, in the language of McFadden et al. (1977)), and $d_{ih}$ is the travel time between the patient's zip code and the hospital. The function $f(d_{ih}, z_{ik}, \alpha h)$ represents distance interacted with patient characteristics and hospital indicators.[8] Thus, the first term includes hospital quality through hospital indicators and allows for heterogenous preferences for hospital quality through interactions between patient characteristics and hospital indicators. The second term allows for horizontal differentiation through distance and allows for heterogeneous preferences over the cost of distance through polynomials of distance interacted with patient characteristics.

In this paper, we focus on one logit model (*Logit*) that includes interactions of hospital indicators with disease acuity, major diagnostic category, and travel time as well as interactions of several patient characteristics – disease acuity, race, sex, age, diagnostic category – with travel time and the square of travel time. This model flexibly accomodates the possibility of heterogeneous preferences over travel time and hospital quality, and has been used

[8]For travel time, we use ArcGIS to calculate the travel time (including traffic) between the centroid of the patient's zip code of residence and each hospital's address.

in recent work on this subject (Garmon, 2017).[9] We estimate this model via maximum likelihood. We use the recovered structural parameters and the new choice set to predict post-disaster choice probabilities.

## 3.2 Machine Learning Models

We now examine two types of machine learning models: a regularization model and a set of decision tree models. These models do not impose the economic assumption that consumers care about the cost of travel time, but allow spatial differentiation in demand by allowing choice probabilities to vary by zip code. Separately, these models are not necessarily unbiased or consistent (Athey and Imbens, 2017).

### 3.2.1 Regularization

In the *Logit* model described above, the researcher decided which covariates to include. A machine learning approach to this same problem is to allow an algorithm to select covariates. We implement a LASSO regression (Tibshirani, 1996) that penalizes the absolute value of coefficients.[10] The parameter estimates recovered by a LASSO model are biased towards 0, and will not generally be consistent (Hastie et al., 2009, p. 91).

To construct the set of possible explanatory variables for our implementation, we interact each of the hospital indicator variables with two way interactions between our set of other predictors. To give an example, one possible explanatory variable would be a specific hospital's indicator variable interacted with a zip code interacted with a MDC code. Constructing variables in this way allows patients from a particular zip code coming into the hospital for a particular condition such as cardiac conditions or pregnancy to have their own valuation of hospital quality.[11]

[9] Raval et al. (2020) shows that this particular econometric model (called *Inter* in that paper) performs better at predicting choices post-disaster compared to several other parametric logit models used in the literature, such as Capps et al. (2003), Ho (2006), Gowrisankaran et al. (2015), and Garmon (2017). We do not examine random coefficients logit models as these have not typically been used in the existing literature on hospital choice, in part because of the availability of individual-level data.

[10] Formally, $-\log L(\beta) + \lambda \sum_{k=1}^{K} |\beta_k|$ where $L(\beta)$ is the log likelihood of a multinomial logit model, $\beta$ are the coefficients of the model, and $\lambda$ is a tuning parameter regulating the degree of shrinkage.

[11] This procedure can generate hundreds or thousands of interactions depending on the dataset. In our implementation, the estimated model provides non-zero weight for about a thousand such interactions for

### 3.2.2 Grouping

An alternative approach to parameterizing $\delta_{ih}$ partitions the space of all patients into a large set of groups, and then assumes homogeneous preferences within each of those groups. Deterministic utility is $\delta_{ih} = \delta_{g(z_i)h}$ for some set of groups $g(z_i)$ that depend upon patient characteristics $z_i$. Thus, this approach is analogous to including an indicator variable for each group-hospital interaction in a multinomial logit model, with the IIA property of proportional substitution holding with each group.

Given a set of groups, predicted choice probabilities can be estimated as the empirical shares of hospitals within each group. For some of the approaches we consider, we use the empirical shares from a single set of groups. For others, we average the shares across different groupings in order to obtain a choice probability for each hospital.

The first grouping model we consider is a semiparametric bin estimator similar to that outlined in Raval et al. (2017) (*Semipar*). For this approach, we place all patients in groups based on their zip code, disease acuity (DRG weight), age group, and area of diagnosis (MDC). Any patient in a group above the minimum group size is assigned choice probabilities based upon the share of patients in that group that go to the various hospitals. For this paper, we use a minimum group size of 20. We then drop a characteristic, reconstruct groups, and again compute group-level shares for the full set of patients, both those previously grouped and those not previously grouped. Because some observations are "re-used," this will lead to relatively smaller variance but higher bias than the approach in Raval et al. (2017), which only used previously non-grouped individuals to compute these choice probabilities. As the size of the dataset goes to infinity, the bias should go to zero as observations will not be recycled.

We drop characteristics in the reverse order listed above (i.e., MDC, age group, etc.) Then, all patients who have not yet been assigned a choice probability and are in groups above the minimum group size are assigned a choice probability based on that round's group-level shares. We continue until all patients are assigned a choice probability or there are no more covariates to group on.[12]

St. Johns and NYU, and about 200 for Moore.

[12]In this last round of grouping, we do not impose a minimum group size restriction. So, for example, if a

While simple to implement and explain, this method for grouping requires the ex ante definition of the order of observable characteristics used for prediction. A set of machine learning models provide algorithmic approaches to allow the data to determine the optimal groups.

The first grouping machine learning model ($DT$) we estimate is a decision tree. While there are many possible ways of estimating tree models, we employ the popular CART approach (Breiman et al., 1984). The CART approach separates the data into two groups at each node based on the split that minimizes the prediction error criterion. Thus, it recursively partitions the data by growing the tree through successive splits. In order to avoid overfitting the data by creating too many splits, the tree model is "pruned" by removing excessive splits that likely contribute little to the out-of-sample performance of the tree. While "pruning" can address the problem of overfitting, the CART approach is known to have a bias towards splitting on covariates with many possible splits (Fu and Simonoff, 2015). However, a single tree will be consistent if the number of nodes is fixed, and the number of observations at each node goes to infinity (Biau et al., 2008, p. 2016).

While a single decision tree is easy to understand and interpret, the literature has tended to conclude that approaches which average the predictions of many tree models have much better predictive power. For example, Breiman (2001b) noted, "While trees rate an A+ on interpretability, they are good, but not great, predictors. Give them, say, a B on prediction." Our second grouping machine learning model ($RF$) leverages this insight, injecting randomness into the tree construction process to create a "random forest" (Breiman, 2001a; Hastie et al., 2009). The random forest introduces two sources of randomness into the formation of trees. First, a whole "forest" of trees are built by estimating different tree models on bootstrap samples of the original dataset. Second, the set of variables that are considered for splitting is different and randomly selected for each tree. To compute choice probabilities for an individual, we average over the group shares relevant to that individual from each of the trees. The consistency of random forests remains an active area of research (Biau et

zip code only has 10 residents, we compute choice probabilities based upon these 10 people. This approach is analagous to estimating different multinomial logit models with group-hospital indicator variables for each level of grouping and assigning choice probabilities to an individual based upon the most refined level of grouping that exceeds the pre-specified minimum group size.

al., 2008; Scornet, 2016), but the literature appears to show that many ways of constructing random forests will be consistent.

Our third grouping machine learning model ($GBM$) also derives from decision tree modeling, but uses "gradient boosting" to generate a multiplicity of trees (Freund and Schapire, 1995; Friedman et al., 2000; Friedman, 2001). Gradient boosting builds off of a single underlying tree structure, creating multiple generations of the original model by overweighting observations that were classified incorrectly in the previous iteration. The final prediction is then a weighted average across all of the different models produced, where a shrinkage parameter scales how much each tree adds to the overall prediction.[13] Biau et al. (2008) shows that the process of averaging may transform the underlying, inconsistent decision trees into a consistent estimator.

## 3.3 Implementation

For all of the machine learning models, we use the same set of predictor variables: the patient's zip code, disease acuity (DRG weight), the Major Diagnosis Category (MDC) of the patient's diagnosis, the patient's age, indicators for medical vs. surgical admission, whether the patient was black, whether the patient was female, and whether the admission was on an emergency basis. We estimate all of the machine learning models using the Scikit package in Python, and set tuning parameters using 3-fold cross-validation. For the three decision tree methods, the main tuning parameter is the minimum size of the node, which we cross validate separately for each experiment, testing values of 10, 25, 50, and 100. For the random forest and gradient boosting methods, we set the number of trees to 5,000. For the regularization model, the main tuning parameter is the shrinkage parameter $\lambda$. All other parameters are set to their default values in Python Scikit. For post-disaster predictions, we estimate probabilities by assuming proportional substitution between the remaining hospitals based on the individual-level probabilities (due to the IIA property of the logit functional form).

[13] In a linear regression, a boosting procedure would overweight observations with large residuals. Boosting can be thought of as an additive expansion in a set of elementary basis functions (in our case, trees).

# 4 Estimation Time

One major consideration when evaluating models is how long they take to run. We provide the computational time required for each of our algorithms in Table 3. These computations were done using the Scikit package in Python 3 on a server where the algorithms were permitted to use up to 40 cores at a time.

We summarize these results for St. Johns, our largest dataset, and Sumter, our smallest; St. Johns has about 14 times the number of admissions as Sumter. The fastest machine learning algorithm is the decision tree; it took 1 second to estimate the decision tree for Sumter and about 14 seconds for St. Johns. Of the machine learning algorithms that generalize decision trees, *RF* is by far the fastest. For Sumter it took 39 seconds, while for St. Johns it took 6 minutes. *GBM* is two orders of magnitude slower than random forest, taking 16 hours for St. Johns and 28 minutes for Sumter.[14] Finally, *Regular* is three orders of magnitude slower than random forest, taking 6 days to run for St. Johns and about 3.4 hours for Sumter.

Table 3: Computational Time by Machine Learning Algorithm

| | Sumter | St Johns | NYU | Moore | Coney | Bellevue |
|---|---|---|---|---|---|---|
| DT | 1 sec | 14 sec | 11 sec | 3 sec | 9 sec | 8 sec |
| RF | 39 sec | 5.9 min | 4.4 min | 43 sec | 1.9 min | 2.1 min |
| GBM | 28.4 min | 16.2 hr | 12.5 hr | 29.6 min | 4.2 hr | 6.8 hr |
| Regular | 3.4 hr | 6.0 days | 3.6 days | 2.1 hr | 22.8 hr | 27.6 hr |

# 5 Average Predictive Performance

We estimate all of the models in Section 3 on training data from the period before the disaster, and assess each model's predictive performance on test data from the period after the disaster. Our measure is the share of choices correctly predicted by the models in the post-disaster test data. We consider a model to predict a choice when the choice probability for that choice is higher than for any of the alternatives. Thus, the logit distributional

[14]The time difference between *GBM* and *RF* is likely because the trees in a random forest model can be constructed in parallel, while the trees in a gradient boosting method are constructed sequentially.

assumption that generates proportional substitution is not required; our prediction criterion is consistent with any error distribution that does not alter the ordering of predicted probabilities.[15]

We first show how well the models perform on an absolute basis and average across the experiments in Figure 1a.[16] We equally weight experiments, not patients. In addition to the benchmark conditional logit, we also compare the machine learning models to a "naive" aggregate share model *Indic*. The *Indic* model assumes that there is no patient heterogeneity, and so patient choices are proportional to observed aggregate shares. Thus, under *Indic*, everyone in the service area is predicted to go to the highest share hospital.

On average, the aggregate share model predicts 28.6% of choices correctly, while our baseline econometric model *Logit* predicts 39.6% of choices correctly. *Semipar*, the semi-parametric bin model, predicts 41.4% of choices correctly. The machine learning models do significantly better – *RF* and *GBM* correctly predict 46.4% of choices, *Regular* 45.6%, and *DT* 44% of choices. Thus, our baseline econometric model predicts 11 percentage points more choices than the aggregate share model, and the best machine learning models 7 percentage points more choices than our baseline econometric model.

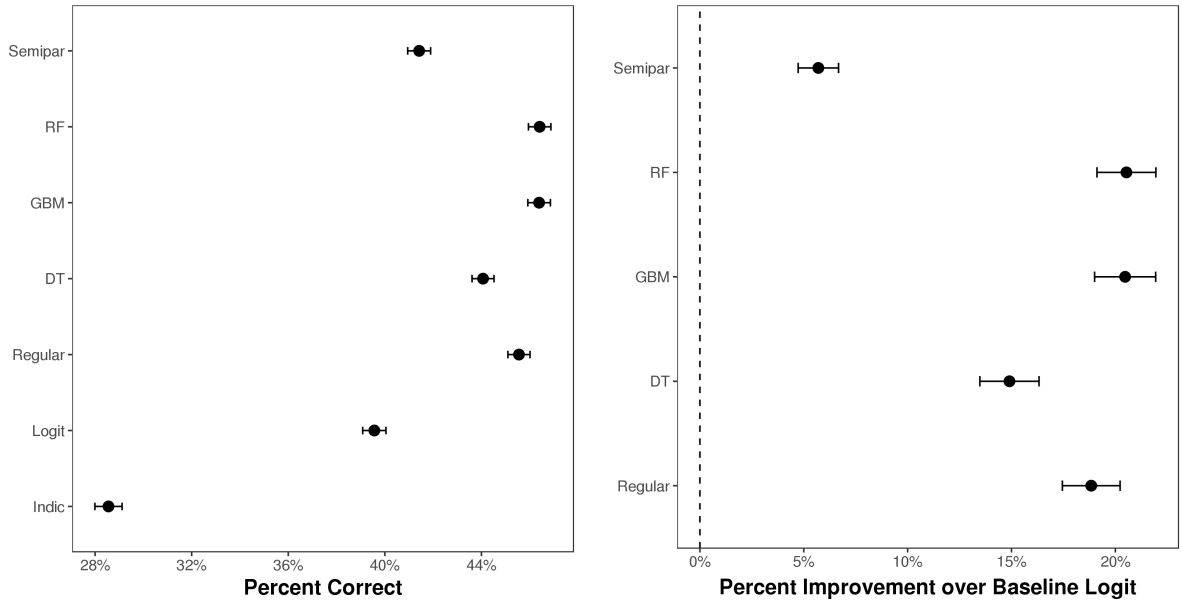
Henceforth, we present the percent improvement in predictive accuracy for all other models relative to the econometric model *Logit*. Figure 1b depicts the percent improvement in the share of correct predictions relative to *Logit*, averaged over all of the experiments. *GBM* and *RF* perform the best, providing a 20.5% increase in predictive accuracy. *Regular* performs 18.8% better than *Logit*, and so is slightly worse than the two best machine learning models. These outperform *DT* and *Semipar*, which are 15% and 6% better than *Logit*. Thus, the two models that build upon an individual decision tree perform the best overall.

We next consider performance at the individual experiment level; in Figure 2, we plot these results, with *RF* in red circles, *GBM* in green triangles, and *Regular* in blue squares.[17]

[15]For the *DT* model, 538 observations for the Sumter experiment have 100% probabilities to the destroyed hospital. For these observations, we set probabilities to the average at the zip code - month level for all other observations. This issue does not affect any other model or *DT* for any other experiment; we discuss this further in the conclusion as one potential problem with machine learning models.

[16]We compute 95% Confidence Intervals based on 500 bootstraps of the test period data, holding the coefficients of the models estimated on the training (pre-disaster) data constant.

[17]We exclude *Semipar* and *DT* for readability; in Table D-3 and Table D-4 in the Appendix we include all of the models.




(a) Absolute (b) Relative to Logit

Figure 1: Predictive Accuracy using Percent Correct – Averaged over all Experiments

**Note:** The left graph is the average percent correct, averaged over all experiments, while the right graph is the average percent correct measured relative to the baseline parametric logit model *Logit*. Bars represent 95% confidence intervals computed from 500 bootstrap replications. See Table D-1 and Table D-2 for tables of the estimates and confidence intervals used to generate these figures.

The models' performance varies substantially across disasters. For example, in Sumter, none of the models perform substantially better than *Logit*, with *RF* the best at a 2.2% improvement. The *DT* and *Semipar* models perform worse than *Logit*. The machine learning models perform dramatically better for Moore, with *RF* and *Regular* having a 63% higher share of correct predictions. For the other four experiments, *RF* and *GBM* perform between 10 to 20% better than *Logit*, and one of the two is the best model. In general, *RF* and *GBM* consistently improve upon the predictive performance of the best of the standard econometric specifications, and are the best two models for 4 of the 6 disasters. Except for Sumter, for which *Regular* performs significantly worse than *Logit*, we cannot statistically reject that the improvement in predictive accuracy is the same for the three best machine learning models.

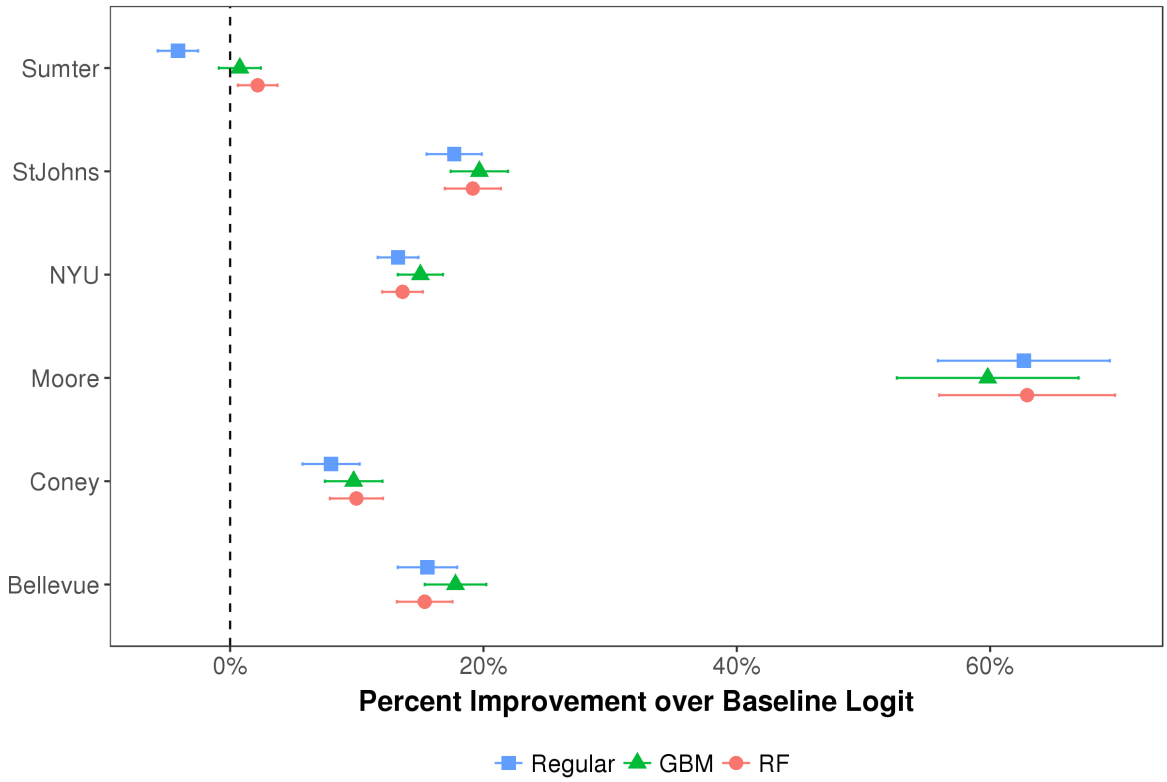



Figure 2: Percent Improvement in Predictive Accuracy using Percent Correct – By Experiment

**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. Bars represent 95% confidence intervals computed from 500 bootstrap replications. See Table D-3 and Table D-4 for tables of the estimates and confidence intervals used to generate this figure.

## 5.1 Validation Sample Performance

In most situations, researchers will not have access to natural experiments like ours in order to assess models, but could gauge performance based on a validation sample that is similar to the training sample. We examine whether performance on a validation sample can provide a good guide to performance after a major change in the choice set by estimating the models on a random 80% sample of the training data (the "train" sample) and then examining their performance on the excluded 20% of the sample (the "validation" sample). We then compare model performance on these samples to our previous results on performance in the "test" sample of post-disaster data in Figure 3.

We find a similar ordering of relative model performance between the validation sample and the test sample. For example, the *GBM* and the *RF* are the two best models using

the training, validation, and test samples. The main exception is the regularization model *Regular*, which appears to overfit less than the decision tree based models. The differences in performance for this model between the training, validation, and test sets are much smaller than for the grouping models; for example, *Regular* performs worse than *DT* in the validation sample but better in the test sample. Overall, our experiments suggest that a validation sample can provide a good guide to model performance even after a major change in environment.

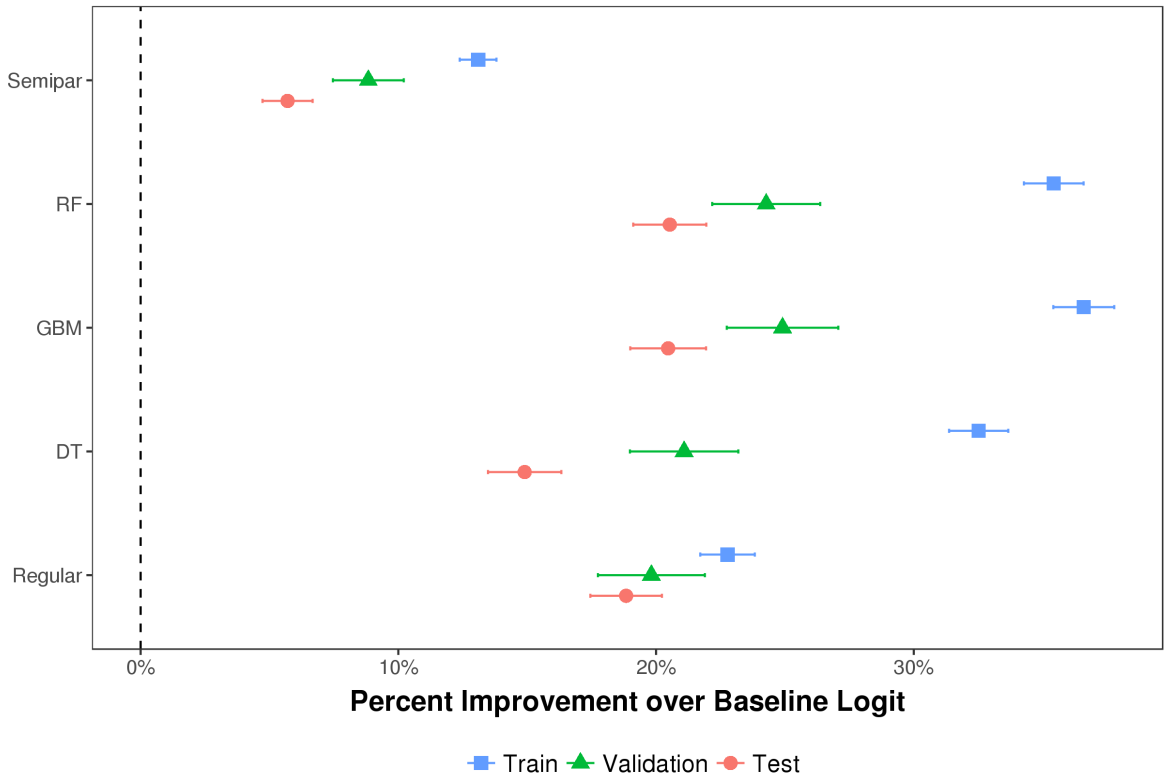



Figure 3: Average Percent Improvement in Predictive Accuracy using Percent Correct, on the Training, Validation, and Test samples

**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. The training sample is a random 80% of the data pre-disaster, the validation sample a random 20% of the data pre-disaster, and the test sample data post-disaster. Bars represent 95% confidence intervals computed from 500 bootstrap replications. See Table D-5 for the table of the estimates and confidence intervals used to generate this figure.

## 5.2 Model Combination

One major finding of machine learning is that ensembles of models can perform better than one individual model (Van der Laan et al., 2007).[18] These findings suggest that combining the predictions from multiple models may lead to better predictions of behavior than using a single "preferred model." In this section, we examine how well a model combination approach does in prediction compared to non-hybrid approaches.

While there are several ways to combine models, we apply a simple regression based approach that has been developed in the literature on optimal model combination for macroeconomic forecasts (Timmermann, 2006). To apply the method to our context, we treat each patient-hospital choice as an observation, and regress each patient's choice of hospital on the predicted probabilities from all of the models in the period after the disaster without including a constant, as below:

$$y_{ih} = \sum_k \beta^k \hat{y}_{ih}^k + \epsilon_{ih}$$

where $y_{ih}$ is the observed choice for patient $i$ and hospital $h$ and $\hat{y}_{ih}^k$ is the predicted probability for patient $i$ and hospital $h$ for model $k$. We constrain the coefficients on the models' predictions to be non-negative and to sum to one. Thus, each coefficient in the regression can be interpreted as a model weight, and models may be given zero weight. We perform this analysis separately for each disaster, which enables us to see the variation in our findings across the different settings.[19]

We examine the performance of the model combination predictions estimated on the 20% validation sample (allowing estimated weights to vary by disaster) in the period after the disaster. The model combination is the best model, but performs only slightly better than *RF* and *GBM*. It provides a 21.4% (95% CI (19.9%, 22.9%)) improvement on *Logit*, compared to 20.5% for *RF* and *GBM*.[20] Given our confidence intervals, the model combination is not

[18] In this study, both *GBM* and *RF* are already combinations of hundreds of base learners and perform very well.

[19] The regression framework implicitly deals with the correlations in predictions across models. If two models are very highly correlated but one is a better predictor than the other, only the better of the two models receives weight in the optimal model combination.

[20] See Table D-8 and Table D-9 for tables of the estimates and confidence intervals used to generate these figures.

statistically differentiable from the best machine learning models. Across experiments, the model combination is the best model for 4 of the 6 experiments, although again by small margins over the best model. Thus, we find evidence that a model combination performs better, but not significantly better, than the best individual model.

## 5.3 Robustness

### 5.3.1 Standard Errors

One might have two major concerns with the reported statistical uncertainty of our results. First, we are implicitly testing several hypotheses at once, such as the performance of several different models or several different models across several disasters. We have thus examined statistical tests that use the Bonferroni correction to adjust for multiple hypotheses. Averaging across experiments, we would reject the null hypothesis that each of the machine learning models performs the same as the *Logit* model. In the results by experiment reported in Figure 2, we would reject that all of the machine learning models perform the same as the *Logit* model for all of the experiments except Sumter; for Sumter, we fail to reject the null for every model.[21] In Section 6, we discuss why we fail to reject the null hypothesis that the algorithmic models predict the same as *Logit* for the Sumter experiment.

Second, in our baseline results, we construct our confidence intervals by estimating the prediction models once, holding the parameter estimates constant, and then bootstrapping model predictions on the test set. This approach would allow for sampling error in the test set, but not modeling error or sampling error in the training data.

We examine the importance of error in the training data by bootstrapping estimates of the model on the training data. We do so for the random forest model because it is quick to estimate, as well as for the logit models, using 200 bootstrap simulations. If we compare each bootstrap model estimate on the same test set, we have a slightly smaller confidence interval for RF relative to baseline. The percent correct for *RF* averaged across experiments has a 95% CI of (46.2%, 46.6%) compared to (45.9%, 46.9%) in our baseline, with an improvement over *Logit* of (19.8%, 21.3%) compared to (19.1%, 22.0%) at baseline.

[21] We also reject the null averaging across all experiments for *Semipar* compared to *Logit*, and reject the null for *Semipar* for each experiment except Sumter and Coney.

If we allow for sampling error in both the training and test set by examining predictions of each bootstrap model estimate on a bootstrapped test set, we obtain the same confidence intervals as our baseline. Thus, we are confident that our results are robust to issues related to the statistical uncertainty of our estimates.

### 5.3.2 Changes in Patient Preferences

Our research design requires that the disaster did not affect the preferences of patients seeking inpatient care. Patients' preferences for a given hospital might have been affected by the disaster if it became substantially more burdensome to travel to a hospital, or because patients were forced to move. For all four disasters, we found that the extent of the damage was limited compared to the size of the affected hospitals' service areas. In Appendix A, we display maps showing the extent of the destruction and summaries of our qualitative research into the timeline of recovery. This gives us confidence that consumers' travel costs to the non-destroyed hospitals did not change after the disaster, after we drop the immediate post-disaster period.

Another concern is that the extensive margin, the absolute number of patients seeking care, falls after the disaster. In Appendix C.3, we show 6% to 14% drops in the number of patients per month in the service area. However, because we condition on patient characteristics, we only require that individual post-disaster patients' preferences are analogous to observably similar pre-disaster ones. The number of patients, or the characteristics of patients on observable dimensions, are allowed to change over time.

Nevertheless, we further attempt to address the possibility of changes in patient preferences in several ways. First, in Appendix C.1 we compare areas with more or less disaster damage for three of the disasters; presumably areas with less damage would be less likely to have patient preferences change. Second, in Appendix C.2, we restrict attention to patients seeking care for more acute conditions such as pregnancy or cardiac problems. Such patients are very unlikely to try to defer seeking care even if their preferred option was destroyed. We also examine differences by the acuity of the diagnosis and the identity of the payer. Our main finding of substantial improvements in predictive performance for machine learning models over our conditional logit baseline continue to hold in these subsamples.

In Appendix C.4, we also show that our findings continue to hold after using RMSE instead of percent correct as a prediction criterion.

# 6 Performance in Changing Environments

The above results demonstrate that, on average, the machine learning models we test tend to predict better than conventional econometric models after the disaster induced change in the choice set. However, this could be because we include many patients for whom their preferred hospital was unaffected by the disaster, and so the destruction of a non-preferred hospital had no effect on their choices.

In this section, we leverage the comparatively unique nature of our context and data to test how well these models predict after a change in the choice environment. In this context, one might worry that a machine learning approach that is fit based on its ability to predict choices in the pre-period may be overfit. That overfit model, with parameters that are not precisely estimated, could lead to very poor counterfactual predictions. In other words, these models ability to provide insights into causal inference would be very poor.

In our implementation, we focus on the patients who were more likely to be substantially affected by the elimination of their preferred hospital following the natural disaster. We do this in three ways. First, we examine predictive accuracy across patients as a function of the probability a similar patient would have gone to the destroyed hospital in the pre-disaster period. We calculate these probabilities based on the groups constructed by *Semipar*. Second, for our New York and California hospitals, we look at patients that had an admission at the destroyed hospital in the pre-period. Third, we examine the weight that the model combination approach places on the different models between the validation and test datasets. For all three, we find that the relative improvement of the machine learning models over the econometric model shrinks for patients more likely to have had a major change in their choice set.

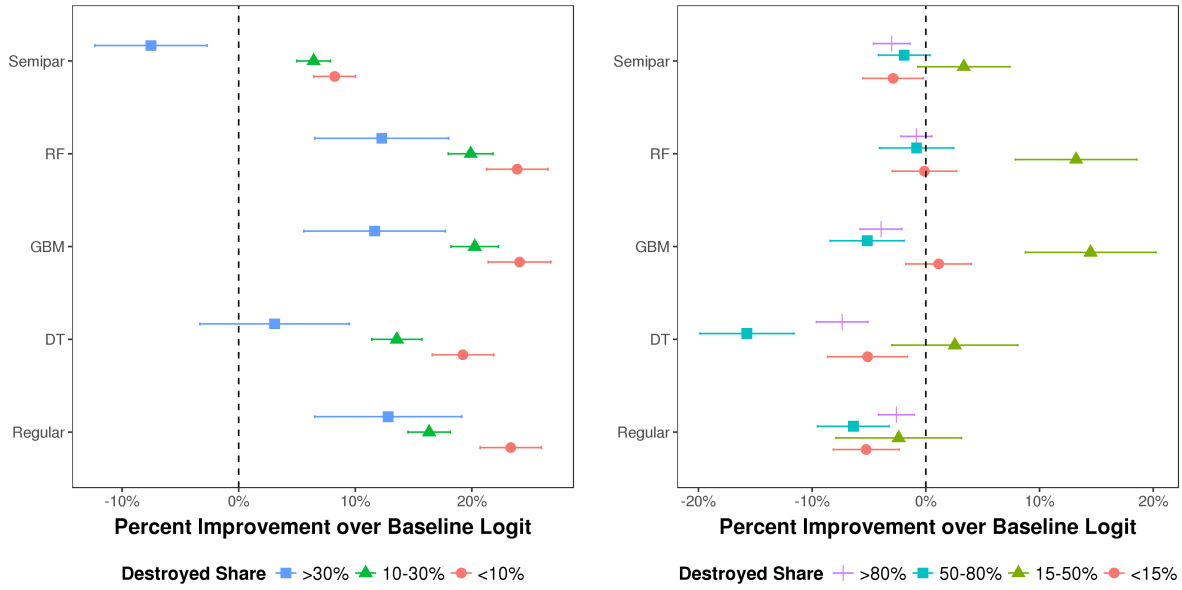
## 6.1 Probability of Using Destroyed Hospital

In Figure 4a, we show the performance of the machine learning models relative to *Logit*, broken down by the share of discharges of the destroyed hospital in the pre-disaster period. The figure shows that the relative improvement of all of the machine learning models over *Logit* is declining in the share of the destroyed hospital. *GBM* and *RF* continue to improve over *Logit*, but they are only 12% better for groups for which the share of the destroyed hospital was above 30%, compared to a 24% improvement for groups for which the share of the destroyed hospital was below 10%. *DT* is only 3% better than *Logit* for groups for which the pre-disaster share of the destroyed hospital is above 30%, while *Semipar* performs worse than *Logit* for these groups. The improvement over our baseline parametric logit for groups with a high share of the destroyed hospital is significantly below the improvement for groups with a low share of the destroyed hospital for the machine learning models tested.

While instructive, only a few patients had a predicted share of the destroyed hospital greater than 30% for many disasters. Therefore, we also look at results separately for Sumter, because the pre-disaster share of the destroyed hospital was 50%, and there was significant variation across groups in the pre-disaster share of the destroyed hospital. We display these results in Figure 4b. In general, machine learning models did worse in areas with a larger share of the destroyed hospital, with all of the models performing worse than *Logit* for a destroyed hospital share of 50% or greater. For example, *RF* is 1% worse than *Logit*, and *GBM* is 4 to 5% worse than *Logit*, for groups with a predicted destroyed hospital share above 50%. In contrast, many of the models perform better than *Logit* for groups with a predicted share of the destroyed hospital between 15 and 50%.

## 6.2 Previous Patients

For our second approach, we focus on predictions for patients with a previous admission in the destroyed hospital. Prior research suggests that these patients are more likely to have gone to the destroyed hospital in the absence of the disaster (Raval and Rosenbaum, 2018). We have a total of 633, 491, 624, and 1,036 admissions for such patients for NYU, Bellevue, Coney, and St. Johns respectively. We depict these results in Figure 5, and compare them




(a) Average (b) Sumter

Figure 4: Percent Improvement in Predictive Accuracy for Percent Correct By Share of Destroyed Hospital

**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. Bars represent 95% confidence intervals computed from 500 bootstrap replications. See Table D-6 and Table D-7 for tables of the estimates and confidence intervals used to generate these figures.

to the average across these four experiments using all patients.

We find that the relative performance of machine learning models falls when only examining previously admitted patients. On average, the *Semipar* and *DT* models are worse than *Logit* on the sample of patients with a previous admission. *RF*, *GBM*, and *Regular* are 6% to 7% better than *Logit* on the previously admitted patients, compared to 13% to 16% better on all patients.

## 6.3 Model Combination Weights

Overall, the previous results suggest that machine learning approaches perform less well relative to standard econometric approaches when focusing on people with the largest change in their choice environment. Using our model combination approach, we now show that the role for traditional parametric demand models increases in such situations. In Table 4, we display the average model weights on the validation sample data, the post-disaster test data,

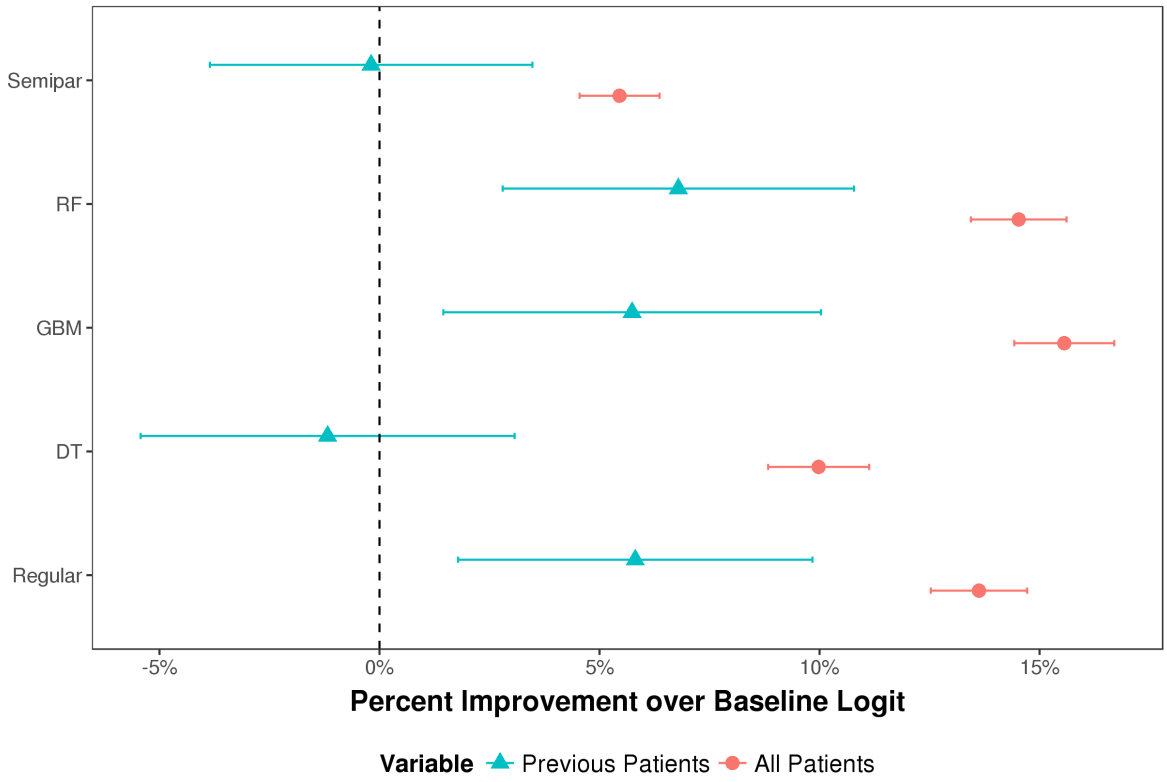



Figure 5: Percent Improvement in Predictive Accuracy for Percent Correct for Previous Patients

**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. Bars represent 95% confidence intervals computed from 500 bootstrap replications. See Table D-10 for the table of the estimates and confidence intervals used to generate this figure.

and only observations in the post-disaster test sample for which the destroyed hospital had at least a 30% share in their *Semipar* group.

We find that the parametric logit model has a much larger share in the post-disaster model combination, especially when the disaster had a large effect on the choice environment. The share of *Logit* rises from 2% using the validation sample, to 9% using the test sample, to 18% using the test sample on only observations with a destroyed hospital share above 30% in their *Semipar* group. In addition, looking just at the results for Sumter, which had the biggest change in choice set, we found that the model combination's share of *Logit* is very large at 48% to 57% for the datasets based on the test data, but only 11% in the validation dataset. Thus, the parametric logit model has a greater role in the model combination with

larger changes in the choice set.[22]

Table 4: Average Model Weights for Optimal Model Combination

| Dataset | Validation | Test | Test, Destroyed Share $\geq 30\%$ |
|---|---|---|---|
| Logit | 0.02 | 0.09 | 0.18 |
| RF | 0.16 | 0.26 | 0.22 |
| GBM | 0.67 | 0.43 | 0.27 |
| DT | 0.15 | 0.06 | 0.09 |
| Regular | 0.00 | 0.15 | 0.20 |
| Semipar | 0.00 | 0.02 | 0.03 |

**Note:** The second through fourth columns provides the average weight for each model across the different experiments using the 20% validation sample, the test sample after the disaster, and only observations in the post-disaster test sample for which the destroyed hospital had at least a 30% share in their *Semipar* group, respectively.

## 6.4 Mechanisms

One potential reason why the machine learning models perform worse relative to the econometric model with a larger change in the choice environment is that a less "local" model is required with a less stable choice environment. That is, the number of data points for each node is cross-validated based on the stable training data, when the choice environment is stable, which may be smaller than optimal in the post period. One reason for this is that the post period probabilities are only based on the choices of patients in the node who did not go to the destroyed hospital, and so may be measured with greater error for nodes where a large share of patients went to the destroyed hospital.

Alternatively, the machine learning models may perform worse because the analysts' domain knowledge – here, encapsulated by travel time as a sufficient statistic for spatial differentiation – is more valuable with larger changes in environment. We test between these two explanations by estimating the random forest model with a small minimum node size (10), and a large minimum node size (100). If the reason for the performance deterioration

[22]The *Regular* model also has its weight increase from 0% in the validation sample to 15% in the test sample to 20% in the test sample on only observations with a destroyed hospital share above 30% in their *Semipar* group. We do not have an explanation for this finding, but it is consistent with the lack of overfitting of the *Regular* model.

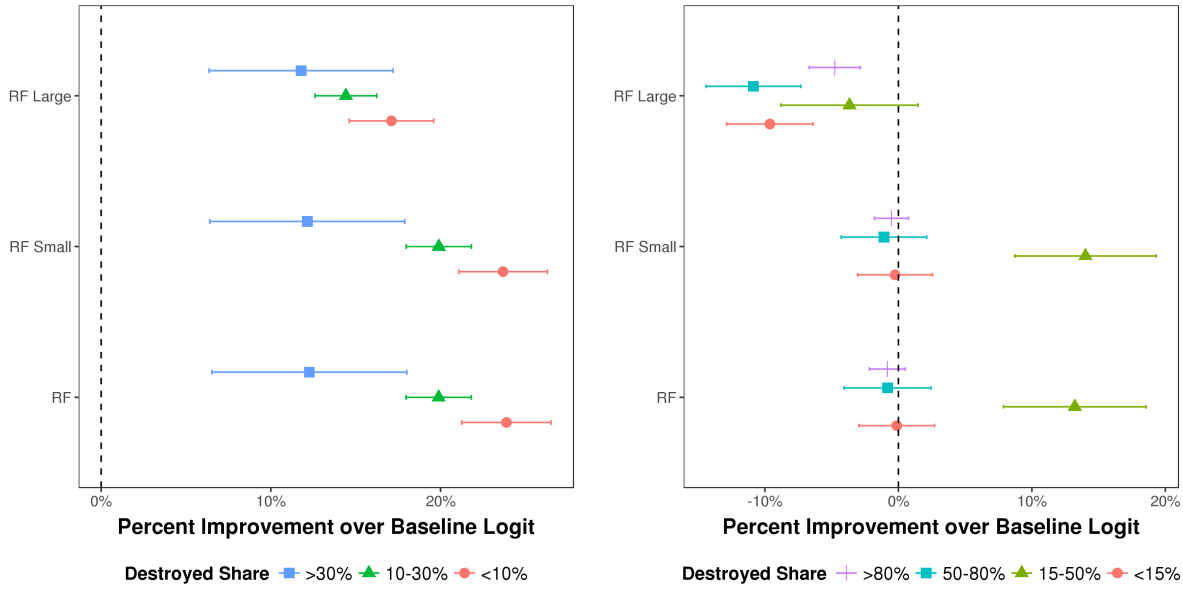
is that a less local, simpler statistical model is needed, the large minimum node size model should perform better in regions with a large change in the choice environment.

We do not find evidence consistent with that prediction. In Figure 6.4, we depict the small minimum node size (*RF Small*) and large minimum node size (*RF Large*) random forest models, together with our previously cross-validated random forest model (*RF*). The left figure depicts performance by pre-period share of the destroyed hospital averaged across all of the experiments, while the right figure examines the Sumter disaster. In both cases, the large node size random forest model performs worse than the small node size random forest model with a small change in the choice set, and equal or worse with a large change in the choice set. All three random forest models are about 12% better than *Logit* for patients with a greater than 30% probability of going to the destroyed hospital. For patients with a less than 10% probability of going to the destroyed hospital, the small node size random forest is 24% better than *Logit*, compared to 17% better for the large node size random forest. For the Sumter disaster, the *RF Small* model always performs better than the *RF Large* model, with the *RF Large* model performing significantly worse than *Logit*. This finding is consistent with there being greater scope for an analyst to bring to bear their domain knowledge when there is a larger change in the choice environment.

# 7 Discussion and Conclusion

In this paper, we show that machine learning models perform significantly better than traditional econometric models in predicting patient decisions following a specific type of treatment. However, we find that their performance relative to econometric models deteriorates as the treatment can be seen as having a larger change on their choice environment. We show that the reason for this deterioration is that using economic domain knowledge to specify parametric structure – in our case, horizontal differentiation due to distance – becomes more important with a changing choice environment.

Therefore, when modeling provider demand, the research or policy question should guide the choice of model. First, researchers often want to make predictions about patient demand without needing to model changes to the choice set. In such cases, a machine learning model




(a) Average (b) Sumter

Figure 6: Percent Improvement in Predictive Accuracy for Percent Correct By Share of Destroyed Hospital, by RF Minimum Node Size

**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. Bars represent 95% confidence intervals computed from 500 bootstrap replications. *RF Large* model is the random forest model estimated using a minimum node size of 100, and the *RF Small* model is the random forest model estimated using a minimum node size of 10. See Table D-11 and Table D-12 for tables of the estimates and confidence intervals used to generate these figures.

will likely provide better predictions. In terms of which machine learning model to choose, we would recommend the random forest model. While we found that leading machine learning models all had very good average predictive performance, the random forest model was the fastest. Since analyses are often time sensitive, there may be large gains from choosing faster approaches. Further, since the random forest model is averaging over many decision trees, the effect of any given tuning parameter on the analysis may be smaller.[23]

In contexts where the researcher needs to model a large change in the choice set, such as hospital merger analysis, we would generally suggest estimating an econometric model – either by itself or in addition to a machine learning model. Our research shows that in changing choice environments, the performance of machine learning models deteriorates. For

[23]We found that in some specifications, the optimal tuning parameter as selected by cross-validation could materially affect the results for the decision tree model. By averaging over many randomized decision trees, this is less likely to be an issue in the random forest and gradient boosting models.

patients with the largest change in their choice environment, the predictive accuracy of the machine learning model is weakly worse than that of the traditional econometric model. The performance of machine learning models may deteriorate even more when the sample size is small, i.e., due to significant subsetting.

In addition, traditional econometric models are designed to examine counterfactuals in which the product characteristics change, such as the effects of entry and product repositioning.[24] Take, for example, the effect of entry of a new hospital; with the econometric model estimated in this paper, one would have to make an assumption on the quality of the new hospital through the fixed effect of the new hospital and add it to the existing choice set. By contrast, for the random forest model, one would have to make assumptions on the quality of the new hospital for each of hundreds of endogenously determined groups and thousands of trees, which would be very complicated to do in practice. Therefore, for the foreseeable future we see machine learning and econometric approaches as being complementary approaches for businesses and policy makers alike.

Overall, we believe that our results illustrate the complementarity of data and domain knowledge. When the machine learning algorithm is being used purely to predict behavior without any change in the surrounding environment, an off the shelf algorithm will likely do a good job at prediction (Mullainathan and Spiess, 2017). However, as the environment changes, a pure prediction model may do a poor job of predicting in the new setting. Therefore, when the choice environment changes, researchers need to use their domain knowledge and incorporate that into their estimating framework in order to better predict future outcomes. This is an important note to keep in mind as researchers develop and use methods that use the tools of machine learning to conduct causal inference (e.g., Belloni et al., 2014; Wager and Athey, 2018).

[24] See Raval and Rosenbaum (forthcoming) and Raval and Rosenbaum (2018) for examples of each in the hospital context.

# References


**Agrawal, Ajay, Joshua Gans, and Avi Goldfarb**, *Prediction Machines: the Simple Economics of Artificial Intelligence*, Harvard Business Press, 2018.

**Athey, Susan**, "Beyond Prediction: Using Big Data for Policy Problems," *Science*, 2017, *355* (6324), 483–485.

_ **and Guido Imbens**, "Machine Learning Methods Economists Should Know About," *arXiv preprint arXiv:1903.10075*, 2019.

_ **and Guido W Imbens**, "The state of applied econometrics: Causality and policy evaluation," *Journal of Economic Perspectives*, 2017, *31* (2), 3–32.

**Bajari, Patrick, Denis Nekipelov, Stephen P Ryan, and Miaoyu Yang**, "Demand Estimation with Machine Learning and Model Combination," Technical Report, National Bureau of Economic Research 2015.

_ **,** _ **,** _ **, and** _ , "Machine Learning Methods for Demand Estimation," *The American Economic Review*, 2015, *105* (5), 481–485.

**Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen**, "Inference on treatment effects after selection among high-dimensional controls," *The Review of Economic Studies*, 2014, *81* (2), 608–650.

**Biau, Gérard, Luc Devroye, and Gäbor Lugosi**, "Consistency of random forests and other averaging classifiers.," *Journal of Machine Learning Research*, 2008, *9* (9).

**Breiman, Leo**, "Random Forests," *Machine Learning*, 2001, *45* (1), 5–32.

_ , "Statistical Modeling: The Two Cultures," *Statistical Science*, 2001, *16* (3), 199–231.

_ **, Jerome Friedman, Charles J. Stone, and R.A. Olshen**, *Classification and Regression Trees*, Chapman and Hall, 1984.

**Capps, Cory, David Dranove, and Mark Satterthwaite**, "Competition and Market Power in Option Demand Markets," *RAND Journal of Economics*, 2003, *34* (4), 737–763.

**Ciliberto, Federico and David Dranove**, "The Effect of Physician–Hospital Affiliations on Hospital Prices in California," *Journal of Health Economics*, 2006, *25* (1), 29–38.

**der Laan, Mark J. Van, Eric C. Polley, and Alan E. Hubbard**, "Super Learner," *Statistical Applications in Genetics and Molecular Biology*, 2007, *6* (1).

**Farrell, Joseph, David J. Balan, Keith Brand, and Brett W. Wendling**, "Economics at the FTC: Hospital Mergers, Authorized Generic Drugs, and Consumer Credit Markets," *Review of Industrial Organization*, 2011, *39* (4), 271–296.

**Freund, Yoav and Robert E. Schapire**, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," in "European Conference on Computational Learning Theory" Springer 1995, pp. 23–37.

**Friedman, Jerome H.**, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, 2001, pp. 1189–1232.

**Friedman, Jerome, Trevor Hastie, and Robert Tibshirani**, "Additive Logistic Regression: a Statistical View of Boosting," *The Annals of Statistics*, 2000, *28* (2), 337–407.

**Fu, Wei and Jeffrey S Simonoff**, "Unbiased regression trees for longitudinal and clustered data," *Computational Statistics & Data Analysis*, 2015, *88*, 53–74.

**Garmon, Christopher**, "The Accuracy of Hospital Merger Screening Methods," *The RAND Journal of Economics*, 2017, *48* (4), 1068–1102.

**Gaynor, Martin, Kate Ho, and Robert J. Town**, "The Industrial Organization of Health-Care Markets," *Journal of Economic Literature*, 2015, *53* (2), 235–284.

**Gaynor, Martin S., Samuel A. Kleiner, and William B. Vogt**, "A Structural Approach to Market Definition with an Application to the Hospital Industry," *The Journal of Industrial Economics*, 2013, *61* (2), 243–289.

**Ghili, Soheil**, "Network formation and bargaining in vertical markets: The case of narrow networks in health insurance," 2016.

**Gowrisankaran, Gautam, Aviv Nevo, and Robert Town**, "Mergers when Prices are Negotiated: Evidence from the Hospital Industry," *American Economic Review*, 2015, *105* (1), 172–203.

**Hastie, Trevor, Robert Tibshirani, and Jerome Friedman**, *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media, 2009.

**Ho, Kate and Robin S. Lee**, "Equilibrium Provider Networks: Bargaining and Exclusion in Health Care Markets," *American Economic Review*, February 2019, *109* (2), 473–522.

**Ho, Katherine**, "The Welfare Effects of Restricted Hospital Choice in the US Medical Care Market," *Journal of Applied Econometrics*, 2006, *21* (7), 1039–1079.

**Hotelling, Harold**, "Stability in Competition," *The Economic Journal*, 1929, *39* (153), 41–57.

**McFadden, Daniel, Antti Talvitie, Stephen Cosslett, Ibrahim Hasan, Michael Johnson, Fred Reid, and Kenneth Train**, *Demand Model Estimation and Validation*, Vol. 5, Institute of Transportation Studies, 1977.

**Mullainathan, Sendhil and Jann Spiess**, "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, 2017, *31* (2), 87–106.

**Petek, Nathan**, "The Marginal Benefit of Inpatient Hospital Treatment: Evidence from Hospital Entries and Exits," *mimeo*, 2016.

**Raval, Devesh and Ted Rosenbaum**, "Why Do Previous Choices Matter for Hospital Demand? Decomposing Switching Costs from Unobserved Preferences," *Review of Economics and Statistics*, 2018, *100* (5), 906–915.

**_ and _** , "Why is Distance Important for Hospital Choice? Separating Home Bias from Transport Costs," *Journal of Industrial Economics*, forthcoming.

_ , _ , **and Nathan E. Wilson**, "Using Disaster Induced Closures to Evaluate Discrete Choice Models of Hospital Demand," *mimeo*, 2020.

_ , _ , **and Steven A. Tenn**, "A Semiparametric Discrete Choice Model: An Application to Hospital Mergers," *Economic Inquiry*, 2017, *55*, 1919–1944.

**Rose, Sherri**, "Machine Learning for Risk Adjustment," *Health Services Research*, 2016, *51* (6), 2358–2374.

**Scornet, Erwan**, "On the asymptotics of random forests," *Journal of Multivariate Analysis*, 2016, *146*, 72–83.

**Tibshirani, Robert**, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996, pp. 267–288.

**Timmermann, Allan**, "Forecast Combinations," *Handbook of Economic Forecasting*, 2006, *1*, 135–196.

**Wager, Stefan and Susan Athey**, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, 2018, *113* (523), 1228–1242.

# A Background on Disasters

In this section, we give brief narrative descriptions of the destruction in the areas surrounding the destroyed hospitals.

## A.1 St. John's (Northridge Earthquake)

On January 17th, 1994, an earthquake rated 6.7 on the Richter scale hit the Los Angeles Metropolitan area 32 km northwest of Los Angeles. This earthquake killed 61 people, injured 9,000, and seriously damaged 30,000 homes. According to the USGS, the neighborhoods worst affected by the earthquake were the San Fernando Valley, Northridge and Sherman Oaks, while the neighborhoods of Fillmore, Glendale, Santa Clarita, Santa Monica, Simi Valley and western and central Los Angeles also suffered significant damage.[25] Over 1,600 housing units in Santa Monica alone were damaged with a total cost of $70 million.[26]

The earthquake damaged a number of major highways of the area; in our service area, the most important was the I-10 (Santa Monica Freeway) that passed through Santa Monica. It reopened on April 11, 1994.[27] By that time, many of those with damaged houses had found new housing.[28]

Santa Monica Hospital, located close to St. John's, remained open but at a reduced capacity of 178 beds compared to 298 beds before the disaster. In July 1995, Santa Monica Hospital merged with UCLA Medical Center.[29] St. John's hospital reopened for inpatient services on October 3, 1994, although with only about half of the employees and inpatient beds and without its North Wing (which was razed).[30]

## A.2 Sumter (Americus Tornado)

On March 1, 2007, a tornado went through the center of the town of Americus, GA, damaging 993 houses and 217 businesses. The tornado also completely destroyed Sumter Regional Hospital. An inspection of the damage map in the text and GIS maps of destroyed structures suggests

[25] See http://earthquake.usgs.gov/earthquakes/states/events/1994_01_17.php.

[26] See http://smdp.com/santa-monicans-remember-northridge-earthquake/131256.

[27] See http://articles.latimes.com/1994-04-06/news/mn-42778_1_santa-monica-freeway.

[28] See http://www.nytimes.com/1994/03/17/us/los-angeles-is-taking-rapid-road-to-recovery.html?pagewanted=all.

[29] See http://articles.latimes.com/1995-07-21/business/fi-26439_1_santa-monica-hospital-medical-center.

[30] See http://articles.latimes.com/1994-09-23/local/me-42084_1_inpatient-services.

that the damage was relatively localized – the northwest part of the city was not damaged, and very few people in the service area outside of the town of Americus were affected.[31] Despite the tornado, employment remains roughly constant in the Americus Micropolitan Statistical Area after the disaster, at 15,628 in February 2007 before the disaster and 15,551 in February 2008 one year later.[32]

While Sumter Regional slowly re-introduced some services such as urgent care, they did not reopen for inpatient admissions until April 1, 2008 in a temporary facility with 76 beds and 71,000 sq ft of space. Sumter Regional subsequently merged with Phoebe Putney Hospital in October 2008, with the full merge completed on July 1, 2009. On December 2011, a new facility was built with 76 beds and 183,000 square feet of space.[33]

## A.3 NYU, Bellevue, and Coney Island (Superstorm Sandy)

Superstorm Sandy hit the New York Metropolitan area on October 28th - 29th, 2012. The storm caused severe localized damage and flooding, shutdown the New York City Subway system, and caused many people in the area to lose electrical power. By November 5th, normal service had been restored on the subways (with minor exceptions).[34] Major bridges reopen on October 30th and NYC schools reopen on November 5th.[35] By November 5th, power is restored to 70 percent of New Yorkers, and to all New Yorkers by November 15th.

FEMA damage inspection data reveals that most of the damage from Sandy occured in areas adjacent to water.[36] Manhattan is relatively unaffected, with even areas next to the water suffering little damage. In the Coney Island area, the island tip suffers more damage, but even here, most block groups suffer less than 50 percent damage. Areas on the Long Island Sound farther east of Coney Island, such as Long Beach, are much more affected.

NYU Langone Medical Center suffered about $1 billion in damage due to Sandy, with its main generators flooded. While some outpatient services reopened in early November, it only partially reopened inpatient services on December 27, 2012, including some surgical services and medical

[31] See https://www.georgiaspatial.org/gasdi/spotlights/americus-tornado for the GIS map.

[32] See http://beta.bls.gov/dataViewer/view/timeseries/LAUMC131114000000005;jsessionid=212BF9673EB816FE50F37957842D1695.tc_instance6.

[33] See https://www.phoebehealth.com/phoebe-sumter-medical-center/phoebe-sumter-medical-center-about-us and http://www.wtvm.com/story/8091056/full-medical-services-return-to-americus-after-opening-of-sumter-regional-east.

[34] See http://web.mta.info/sandy/timeline.htm.

[35] See http://www.cnn.com/2013/07/13/world/americas/hurricane-sandy-fast-facts/.

[36] See the damage map at https://www.huduser.gov/maps/map_sandy_blockgroup.html.

and surgical intensive care. The maternity unit and pediatrics reopened on January 14th, 2013. [37] While NYU Langone opened an urgent care center on January 17, 2013, a true emergency room did not open until April 24, 2014, more than a year later.[38]

Bellevue Hospital Center reopened limited outpatient services on November 19th, 2012.[39] However, Bellevue did not fully reopen inpatient services until February 7th, 2013.[40] Coney Island Hospital opened an urgent care center by December 3, 2012, but patients were not admitted inpatient. It had reopened ambulance service and most of its inpatient beds by February 20th, 2013, although at that time trauma care and labor and delivery remained closed. The labor and delivery unit did not reopen until June 13th, 2013.[41]

## A.4 Moore (Moore Tornado)

A tornado went through the Oklahoma City suburb of Moore on May 20, 2013. The tornado destroyed two schools and more than 1,000 buildings (damaging more than 1,200 more) in the area of Moore and killed 24 people. Interstate 35 was briefly closed for a few hours due to the storm.[42] Maps of the tornado's path demonstrate that while some areas were severely damaged, most nearby areas were relatively unaffected.[43]

Emergency services, but not inpatient admissions, temporarily reopened at Moore Medical Center on December 2, 2013. Groundbreaking for a new hospital took place on May 20, 2014, while the new hospital opened May 6, 2016.[44]

[37] See http://www.cbsnews.com/news/nyu-langone-medical-center-partially-reopens-after-sandy/.

[38] See http://fox6now.com/2013/01/17/nyu-medical-center-reopens-following-superstorm-sandy/ and http://www.nytimes.com/2014/04/25/nyregion/nyu-langone-reopens-emergency-room-that-was-closed-by-hurricane-sandy.html.

[39] See http://www.cbsnews.com/news/bellevue-hospital-in-nyc-partially-reopens/.

[40] See http://www.nbcnewyork.com/news/local/Bellevue-Hospital-Reopens-Sandy-Storm-East-River-Closure-190298001.html.

[41] See http://www.sheepsheadbites.com/2012/12/coney-island-hospital-reopens-urgent-care-center/, http://www.sheepsheadbites.com/2013/02/coney-island-hospital-reopens-er-limited-911-intake/, and http://www.sheepsheadbites.com/2013/06/photo-first-post-sandy-babies-delivered-at-coney-island-hospital-after-labor-and-delivery-unit-reopens/.

[42] See http://www.news9.com/story/22301266/massive-tornado-kills-at-least-51-in-moore-hits-elementary-school.

[43] See http://www.srh.noaa.gov/oun/?n=events-20130520 and http://www.nytimes.com/interactive/2013/05/20/us/oklahoma-tornado-map.html for maps of the tornado's path.

[44] See https://www.normanregional.com/en/locations.html?location_list=2, http://kfor.com/2013/11/20/moore-medical-center-destroyed-in-tornado-to-reopen-in-december/, and https://oklahoman.com/article/5494931/norman-regional-moore-readies-for-reopening-three-

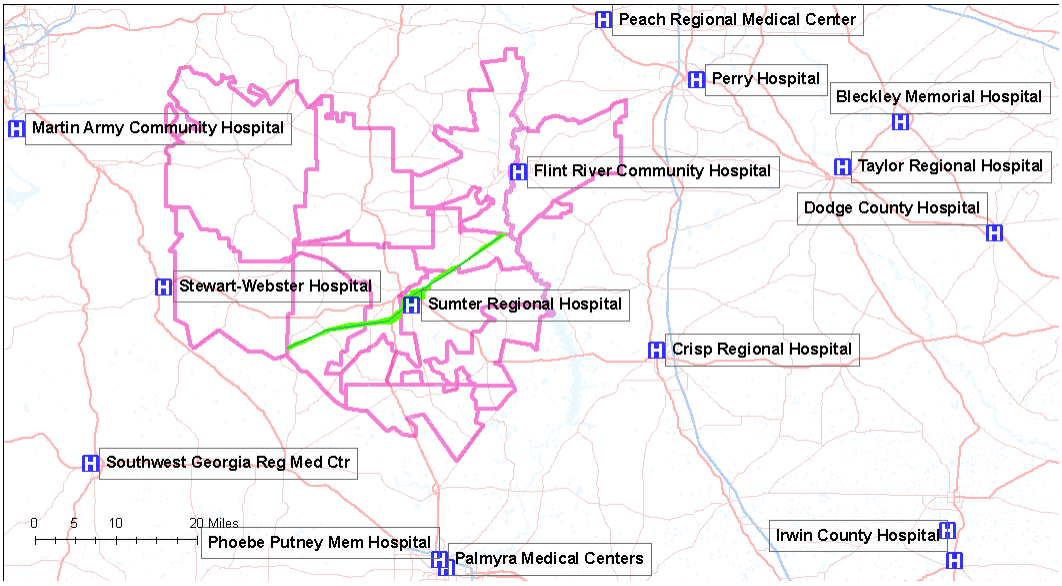



Figure 7: Damage Map in Americus, GA

**Note:** The green line indicates the path of the tornado and the shaded area around it is the government designated damage area. The zip codes included in the service area are outlined in pink. Sources: City of Americus, GA Discharge Data.

## A.5 Geographic Extent of Damage

In this subsection, we present graphical evidence of the scope of damage in Sumter, Moore, New York (NYU, Bellevue, and Coney Island), and Los Angeles in Figure 7 - Figure 10. In each figure, zip codes in the service area are outlined.

Figure 7 shows the path of the tornado that destroyed Sumter Regional Hospital as a green line. The figure indicates that it cut through Americus city without affecting the surrounding areas. As shown in Figure 8, the Moore tornado had a similar effect for the city of Moore relative to its neighboring suburbs.

Figure 9 shows the damage caused by Superstorm Sandy to the areas surrounding NYU, Bellevue, and Coney Island. Flooding – the damage from which is depicted in green shading – primarily affected areas adjacent to water. The actual damage in Manhattan from Sandy – most of which classified by FEMA as "minor" – was concentrated in a relatively small part of the Manhattan hospitals' service areas. On Coney Island, most of the flooding affected the three zip codes at the bottom of the service area that are directly adjacent to Long Island Sound.

Finally, Figure 10 shows the damage in the Los Angeles area from the Northridge earthquake.

`years-after-tornado-ripped-through-hospital?`.

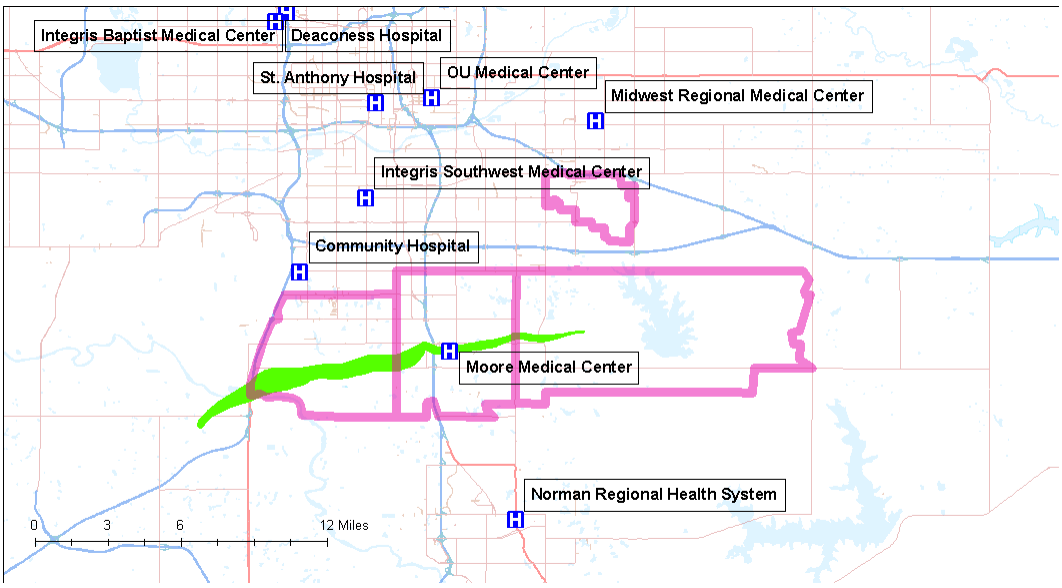



Figure 8: Damage Map in Moore, OK

**Note:** The green area indicates the damage path of the tornado. The zip codes included in the service area are outlined in pink. Sources: NOAA, OK Discharge Data

We depict the intensity of earthquake shaking with darker green shading, and the figure shows that damage was more widespread than in the other disasters. However, while the Santa Monica area was particularly hard hit, many areas nearby suffered comparatively little structural damage from the earthquake.

# B Dataset Construction

For each dataset, we drop newborns, transfers, and court-ordered admissions. Newborns do not decide which hospital to be born in (admissions of their mothers, who do, are included in the dataset); similarly, government officials or physicians, and not patients, may decide hospitals for court-ordered admissions and transfers. We drop diseases of the eye, psychological diseases, and rehabilitation based on Major Diagnostic Category (MDC) codes, as patients with these diseases may have other options for treatment beyond general hospitals. We also drop patients whose MDC code is uncategorized (0), and neo-natal patients above age one. We also exclude patients who are missing gender or an indicator for whether the admission is for a Medical Diagnosis Related Group (DRG). We also remove patients not going to General Acute Care hospitals.

For each disaster, we estimate models on the pre-period prior to the disaster and then validate

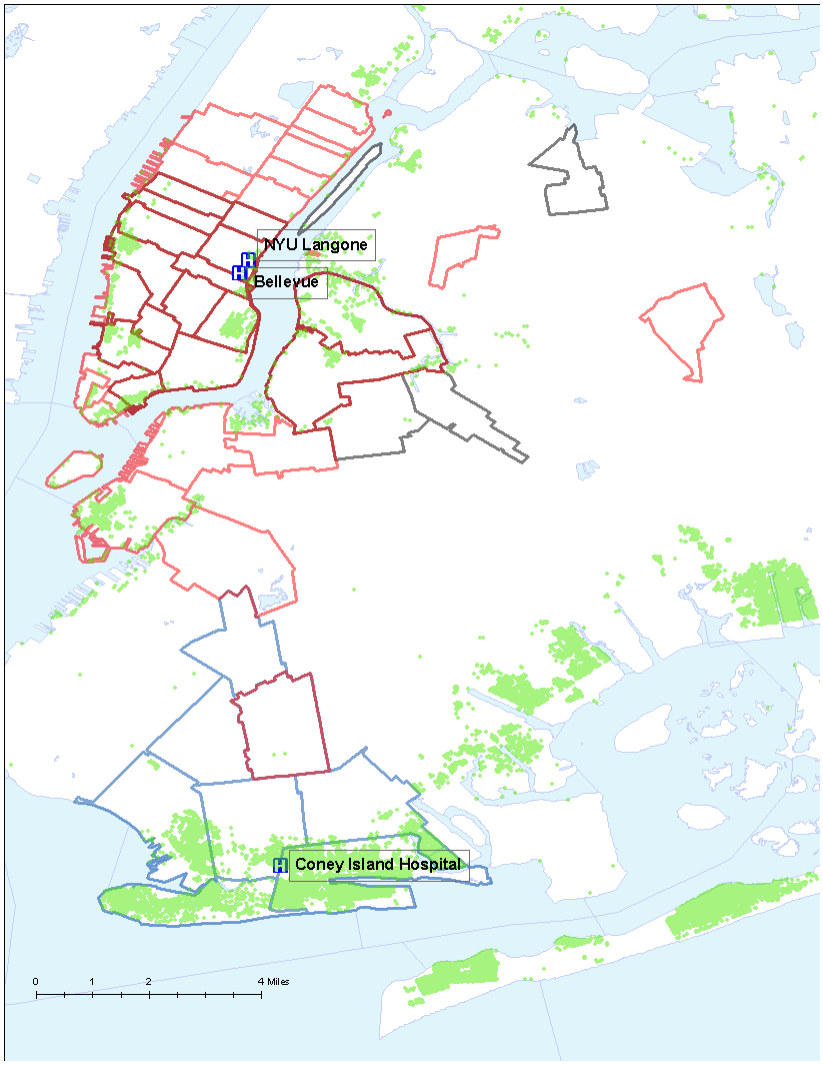



Figure 9: Damage Map in New York, NY

**Note:** Green dots indicate buildings with damage classified as "Minor", "Major", or "Destroyed" by FEMA. The zip codes included in the service area for Bellevue are outlined in gray, for NYU are outlined in pink, and for Coney Island are outlined in blue. The other border colors are for zip codes that are in the service areas of multiple hospitals (maroon is for NYU and Bellevue and red is for NYU and Coney Island). Sources: FEMA, NY Discharge Data

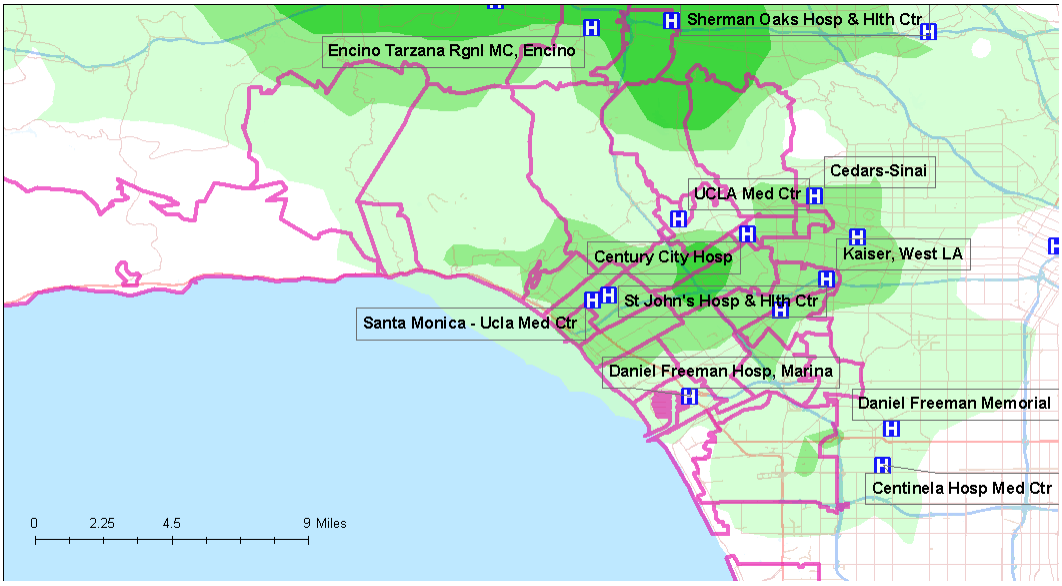



Figure 10: Damage Map in Los Angeles, CA

**Note:** Darker green areas indicate the earthquake intensity measured by the Modified Mercalli Intensity (MMI); an MMI value of 7 reflects non-structural damage and a value of 8 moderate structural damage. The areas that experienced the quake with greater intensity were shaded in a darker color, with the MMI in the area ranging from 7-8.6. Any areas with an MMI of below 7 were not colored. The zip codes included in the service area are outlined in pink. Sources: USGS Shakemap, OSHPD Discharge Data

them on the period after the disaster. In all cases, we omit the month of the disaster from either period, excluding anyone either admitted or discharged in the disaster month. We also omit additional months if our information suggests that the area has not recovered yet. The length of the pre-period and post-period in general also depend upon the length of the discharge data that we have available. Table B-1 contains the disaster date and the pre-period and post-period for each disaster, where months are defined by time of admission.

NYU hospital began limited inpatient service on December 27, 2012; unfortunately, we only have month and not date of admission and so cannot remove all patients admitted after December 27th. Right now, we drop 65 patients admitted in December to NYU; this patient population is very small compared to the size and typical capacity of NYU.

For California, we exclude all patients going to Kaiser hospitals, as Kaiser is a vertically integrated insurer and almost all patients with Kaiser insurance go to Kaiser hospitals, and very few patients without Kaiser insurance go to Kaiser hospitals. This is in line with the literature examining hospital choice in California including Capps et al. (2003). We also exclude February though April 1994, as the I-10 Santa Monica freeway that goes through Santa Monica only reopens

in April.

Table B-1: Pre and Post Periods for Disasters

| Hospital | Closure Date | Pre-Period | Post-Period | Partial Reopen | Full Reopen |
|---|---|---|---|---|---|
| St. Johns | 1/17/94 | 1/92 to 1/94 | 5/94 to 9/94 | 10/3/94 | 10/3/94 |
| Sumter | 3/1/07 | 1/06 to 2/07 | 4/07 to 3/08 | 4/1/08 | 4/1/08 |
| NYU | 10/29/12 | 1/12 to 9/12 | 11/12 to 12/12 | 12/27/12 | 4/24/14 |
| Bellevue | 10/31/12 | 1/12 to 9/12 | 11/12 to 12/12 | 2/7/13 | 2/7/13 |
| Coney | 10/29/12 | 1/12 to 9/12 | 11/12 to 12/12 | 2/20/13 | 6/11/13 |
| Moore | 5/20/13 | 1/12 to 4/13 | 6/13 to 12/13 | 5/7/16 | 5/7/16 |

# C Robustness

In this section, we evaluate the robustness of our conclusions to removing areas with more damage from the disaster, to examining only specific patient groups, and to using RMSE instead of percent correct as a prediction criterion. We find that doing so does not lead to substantially different conclusions than described earlier. We also examine how the case-mix of the service area changes post-disaster, and find evidence of reductions in the number of inpatient admissions after the disaster.

## C.1 Removing Destroyed Areas

Our first approach to evaluating the robustness of our conclusions is to consider the effect of removing the areas most affected by the disaster from our estimates of model performance after the disaster. If destruction from the disaster affects how patients make decisions beyond just the change in the choice set (for example, they are forced to move), then models estimated before the disaster may not be able to predict patients' decisions after the disaster. We focus on Sumter, Coney Island, and Northridge. We do not remove any areas for NYU or Bellevue, as the area immediately nearby these hospitals had very little post-Sandy damage. For Moore, removing the zip codes through which the tornado traversed would remove almost all of the patients from the choice set, so we do not conduct this robustness check for Moore.

For Sumter, we remove the two zip codes comprising the city of Americus, for the tornado mainly damaged the city rather than its outlying areas. For Coney Island, we remove the three

zip codes that submitted the most post-disaster claims to FEMA; these zip codes are on the Long Island Sound and likely suffered more from flooding after Sandy. For St. Johns, we remove zip codes with an average Modified Mercalli Intensity (MMI) of 8 or above based on zip code level data from an official report on the Northridge disaster for the state of California. The US Geological Survey defines MMI values of 8 and above as causing structural damage. This removes 9 zip codes, including all 5 zip codes in Santa Monica.[45] The areas removed tend to have higher market shares for the destroyed hospital. Thus, removing destroyed areas cuts Sumter's market share from about 50 percent to 31 percent, St. John's market share falls from 17 to 14 percent, and Coney's from about 18 to 10 percent.

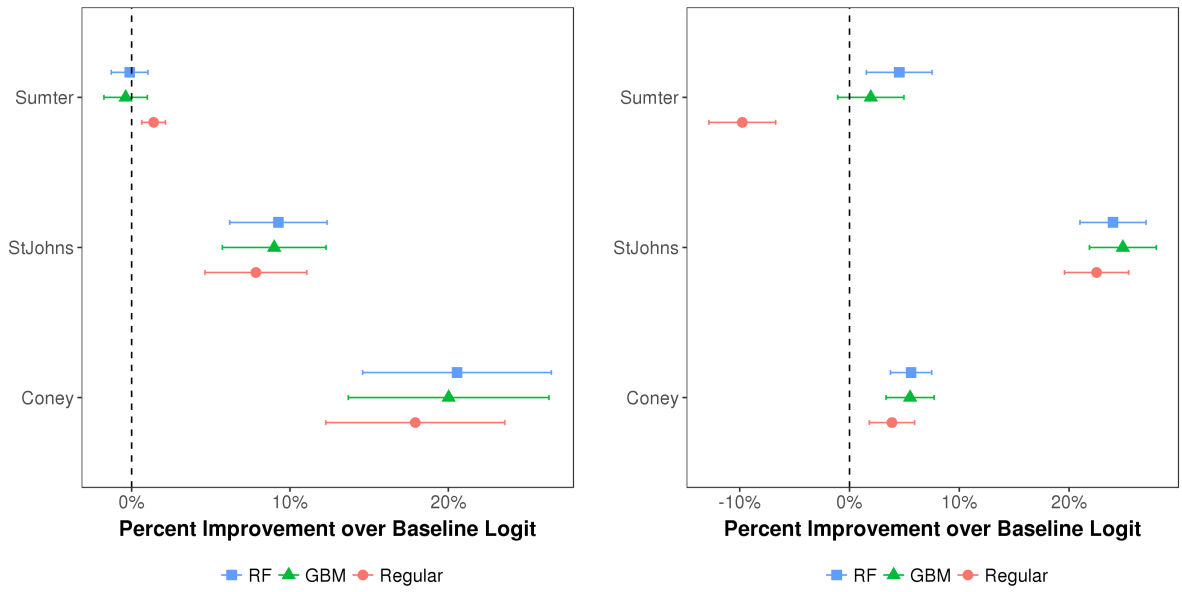
We estimate the models on the full pre-disaster sample but separately evaluate our performance validation measures based on whether the patient came from an area with or without significant damage. In Figure 11, we display these results for the damaged areas in the left figure, and for the relatively non-damaged areas in the right figure. We find that the machine learning models almost always outperform the econometric models for Coney and St. Johns in both the damaged and non-damaged areas, although their margin of improvement is larger in the destroyed areas for Coney and smaller for St. Johns.

For Sumter, *GBM* and *RF* slightly underperform *Logit* in the damaged areas, and out perform *Logit* in the non destroyed areas, consistent with our evidence on how the models performed with different shares of the destroyed hospital in Section 6. *Regular* and *Semipar* perform slightly better than *Logit* in the destroyed areas, but much worse in the non destroyed areas.

## C.2 Patient Heterogeneity

For our second robustness check, we consider the performance of different predictive models for different types for patients. First, we examine seven important classes of patients based on their diagnosis: cardiac patients (with a Major Diagnostic Category of 5), obstetrics patients (with a Major Diagnostic Category of 14), non-emergency as well as emergency patients, and terciles of the disease acuity of patients, measuring disease acuity by DRG weight. We estimate the models on all patients, but then separately examine their performance for patients in the given groups.

[45] The zip codes removed are 31719 and 31709 for Sumter; 90025, 90064, 90401, 90402, 90403, 90404, 90405, 91403, and 91436 for St. Johns; and 11224, 11235, and 11229 for Coney. See http://www.arcgis.com/home/webmap/viewer.html?webmap=f27a0d274df34a77986f6e38deba2035 for Census block level estimates of Sandy damage based on FEMA reports. See ftp://ftp.ecn.purdue.edu/ayhan/Aditya/Northridge94/OES%20Reports/NR%20EQ%20Report_Part%20A.pdf, Appendix C, for the Northridge MMI data by zip code.




(a) Damaged Areas

(b) Non-Damaged Areas

Figure 11: Relative Improvement in Percent Correct – Damaged vs. Non-Damaged Areas

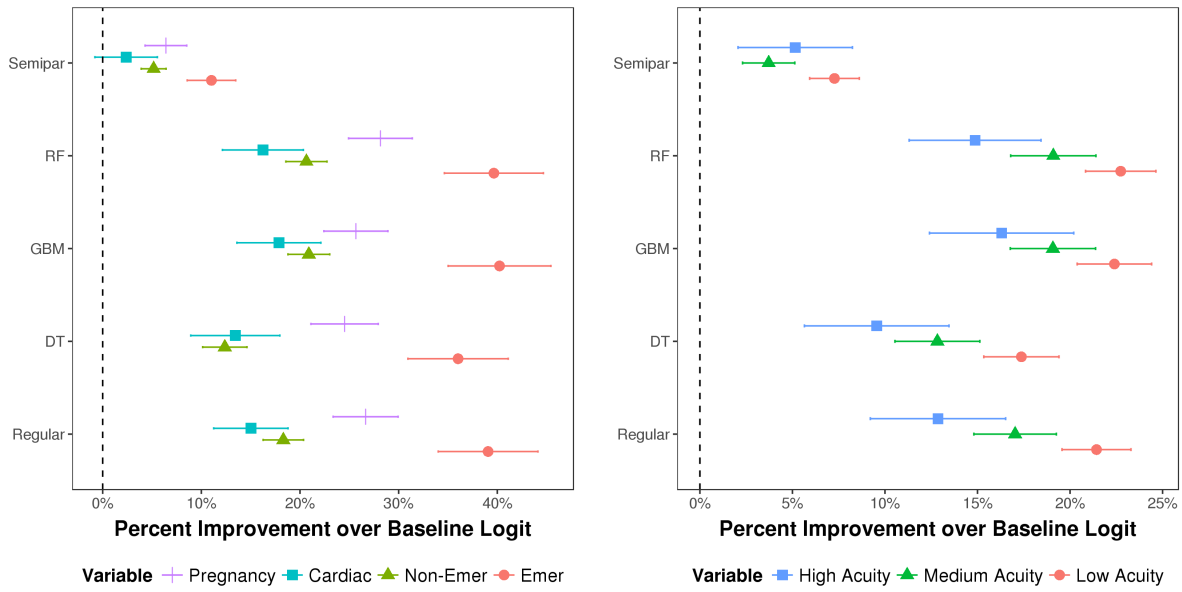
**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. Bars represent 95% confidence intervals computed from 500 bootstrap replications. See Table D-13 and Table D-14 for tables of the estimates and confidence intervals used to generate these figures.

Figure 12 displays the results of our different robustness checks. The machine learning models continue to do significantly better than *Logit* for all of the groups. Their relative performance is better for emergency compared to non-emergency patients, and for pregnancy compared to cardiac patients. For example, *RF* is 44% better than *Logit* for emergency patients compared to 35% better for non-emergency patients, and is 35% better for pregnancy patients compared to 24% better for cardiac patients. We find better relative performance for machine learning models for low acuity patients than medium acuity patients, and medium acuity patients compared to high acuity patients.

In addition, we check whether our conclusions hold if we restrict the data sample to the patient population for different payers. For the Medicare sample, we also reestimate the models on only the Medicare sample, as this sample should have unrestricted access to all the hospitals in the choice set.[46]

We depict these estimates in Figure 13; the machine learning models continue to improve

[46]For the states for which Fee for Service Medicare and Managed Care Medicare are distinguished, we exclude Managed Care Medicare patients. The Medicare sample should have unrestricted access across all of the hospitals in the choice set.




(a) Emergency and MDC

(b) Acuity

Figure 12: Average Relative Improvement in Percent Correct: Robustness

**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. We examine cardiac, pregnancy, emergency, and non-emergency patients separately in the left figure, and disease acuity (DRG weight) divided into terciles in the right figure. Bars represent 95% confidence intervals computed from 500 bootstrap replications. See Table D-15 and Table D-16 for tables of the estimates and confidence intervals used to generate these figures.

over our baseline econometric model for all types of payers. For example, on average, *RF* is 31% better than *Logit* for commercial patients, 19% better for Medicare patients, 15.5% better for Medicare patients (re-estimating the models on Medicare patients only), and 51% better for Medicaid patients. While the machine learning models tend to do relatively better on Medicaid patients compared to commercial patients, and commercial patients compared to Medicare patients, they outperform our baseline parametric logit model for all types of patients.

## C.3 Case Mix

In this section, we examine how the case mix changed from the period before the disaster to the period after the disaster. The case mix could have changed for a couple of reasons. First, patients could have left the service area after the disaster, perhaps because their homes or workplaces were damaged. Second, some patients could have decided not to receive medical assistance after the hospital closest to them was destroyed. Changes in case mix could indicate substantial changes to the service area that make the disaster less of an clean experiment.

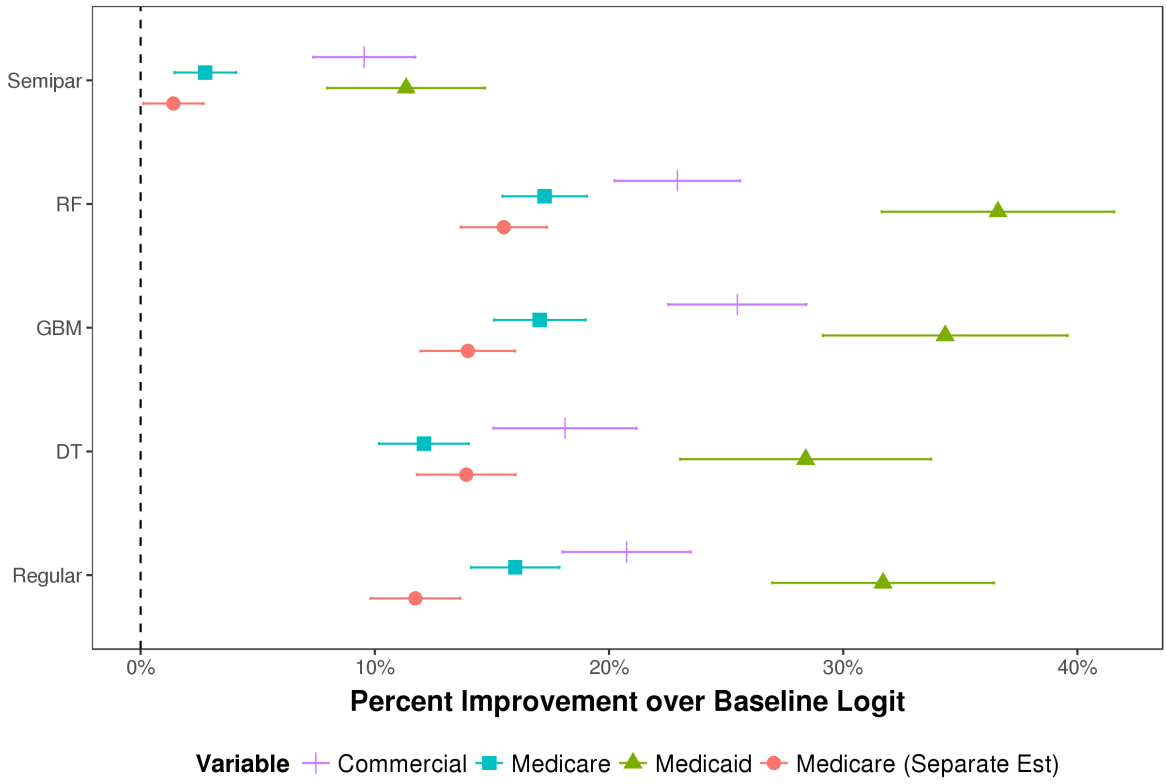



Figure 13: Average Relative Improvement in Percent Correct: Payer Type

**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. For the Medicare (Separate Est) bars, we examine Medicare patients only and reestimate all of the models on the Medicare only sample in order to develop predictions. Bars represent 95% confidence intervals computed from 500 bootstrap replications. See Table D-17 for the table of the estimates and confidence intervals used to generate this figure.

In Table B-2 to Table B-7, we examine changes in case mix across a set of variables including age, fraction aged less than 18, fraction aged above 64, diagnosis acuity (DRG weight), fraction circulatory diagnosis (MDC 5), fraction labor/pregnancy diagnosis (MDC 14), fraction using a commercial payer, fraction using Medicare, and average number of admissions per month. We report the average of each variable in the pre-period, post-period, as well as the percent difference between the two.

There are no large changes in age across the hospitals, except that the fraction admitted under 18 falls by 23 percent for Moore and 45 percent for Sumter. Diagnosis acuity does not change much after the disasters. The only large change in type of insurance is for Sumter, where the fraction of commercial insurance falls by about 30 percent after the disaster. We examined this change; the fraction of patients reporting "Unspecified Other" payer rises precipitously in the first quarter

after the disaster, and then falls back to a small fraction of patients. Our belief is that this reflects improper coding post-disaster.

The number of admissions per month falls in all service areas, ranging from 6 to 8 percent for NYU, Coney, Moore, and St. John's, 11 percent for Bellevue, and 14 percent for Sumter. This likely reflects some extensive margin in inpatient admissions, consistent with the findings of Petek (2016) from hospital exits. The fraction of labor/pregnancy diagnosis rises in all service areas, and by more than 10 percent for Bellevue and Coney, which may be because pregnancies cannot be postponed or ignored and so have no extensive margin. Overall, we do not find major changes in case mix after the disaster, except for the fall in admissions across the service areas and the fall in the under 18 share for Sumter and Moore.

Table B-2: Changes in Case-Mix for Moore

| Variable | Training | Test | Percent Difference |
|---|---|---|---|
| Age | 51.68 | 51.79 | 0.21% |
| Age $< 18$ | 0.06 | 0.05 | -23.37% |
| Age $> 64$ | 0.36 | 0.35 | -2.67% |
| Diagnosis Acuity | 1.41 | 1.44 | 2.23% |
| Circulatory Diagnosis | 0.12 | 0.10 | -12.02% |
| Labor/Pregnancy Diagnosis | 0.20 | 0.22 | 6.86% |
| Commercial Payer | 0.35 | 0.36 | 3.76% |
| Medicare Payer | 0.40 | 0.39 | -1.79% |
| Admissions Per Month | 610 | 560 | -8.22% |

**Note:** The second column is the average of the variable in the pre-disaster training data, while the third column is the average of the variable in the post-disaster test data. The fourth column is the percent difference from the pre-disaster training data to the post-disaster test data.

## C.4 RMSE as Prediction Criterion

Our baseline prediction criterion of percent correct ignores the models' estimates of probabilities for non-selected choices. However, estimates of welfare depend on probabilities of all hospitals in the choice set, and not just the chosen hospital. Therefore, we also present many of our results using root mean squared error across the probabilities of all choices, which penalizes models that incorrectly predict probabilities for low probability hospitals that none of the models would select as the most likely choice. We find similar results to our earlier results with percent correct using RMSE as a prediction criterion.

Table B-3: Changes in Case-Mix for Coney

| Variable | Training | Test | Percent Difference |
|---|---|---|---|
| Age | 57.59 | 57.65 | 0.11% |
| Age $< 18$ | 0.05 | 0.05 | 3.18% |
| Age $> 64$ | 0.46 | 0.47 | 3.05% |
| Diagnosis Acuity | 1.34 | 1.39 | 3.41% |
| Circulatory Diagnosis | 0.20 | 0.19 | -5.17% |
| Labor/Pregnancy Diagnosis | 0.16 | 0.18 | 13.30% |
| Commercial Payer | 0.19 | 0.18 | -6.23% |
| Medicare Payer | 0.46 | 0.47 | 2.26% |
| Admissions Per Month | 5176 | 4833 | -6.63% |

**Note:** The second column is the average of the variable in the pre-disaster training data, while the third column is the average of the variable in the post-disaster test data. The fourth column is the percent difference from the pre-disaster training data to the post-disaster test data.

In Figure 14, we depict the percent improvement in RMSE (so the negative in the change in RMSE) relative to *Logit*, averaged over all of the experiments. We again find that *GBM* and *RF* are the best models. However, the margin of improvement over *Logit* is much smaller; *GBM* and *RF* are both about 3.7% better than the baseline parametric logit model *Logit*. The regularization model *Regular* is about 2.5% better than *Logit*. The only major difference compared to our results for percent predicted is that the *DT* model performs relatively much worse, at only a 0.6% improvement over *Logit*.

In Figure 15, we show these results by disaster. For Sumter, all of the machine learning models are now worse than *Logit*; *RF* is 0.7% worse and *GBM* is 1.5% worse. However, for all of the other experiments, we find that *RF* or *GBM* are the best two models.

# D Supplemental Tables

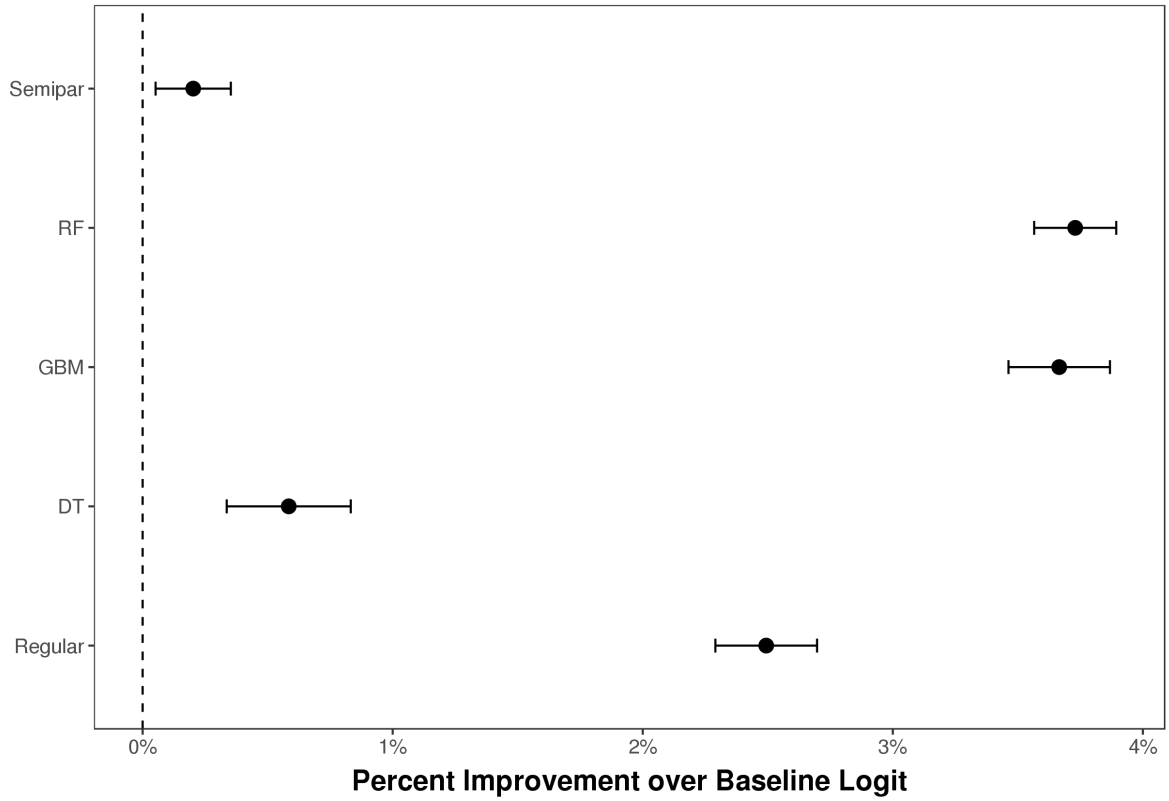



Figure 14: Percent Improvement in Predictive Accuracy using RMSE – Averaged over all Experiments

**Note:** Predictive Accuracy is Measured as RMSE, averaged over all experiments and measured relative to the baseline parametric logit model *Logit*; since we depict percent improvement, the negative is the change in RMSE. Bars represent 95% confidence intervals computed from 500 bootstrap replications. See Table D-18 for a table of the estimates and confidence intervals used to generate this figure.

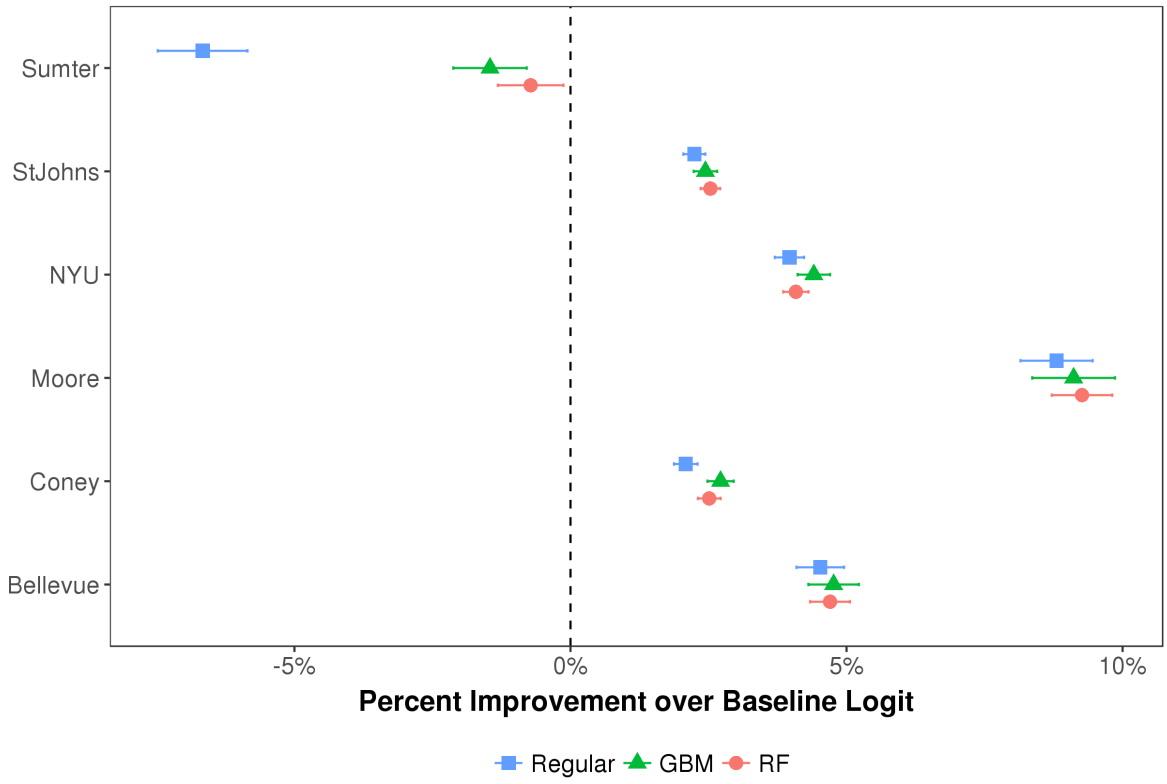



Figure 15: Percent Improvement in Predictive Accuracy using RMSE – By Experiment

**Note:** RMSE measured relative to the baseline parametric logit model *Logit*; since we depict percent improvement, the negative is the change in RMSE. Bars represent 95% confidence intervals computed from 500 bootstrap replications. See Table D-19 for a table of the estimates and confidence intervals used to generate this figure.

Table B-4: Changes in Case-Mix for NYU

| Variable | Training | Test | Percent Difference |
|---|---|---|---|
| Age | 56.09 | 56.61 | 0.93% |
| Age $< 18$ | 0.05 | 0.05 | 2.54% |
| Age $> 64$ | 0.42 | 0.44 | 4.71% |
| Diagnosis Acuity | 1.28 | 1.30 | 1.01% |
| Circulatory Diagnosis | 0.17 | 0.16 | -7.74% |
| Labor/Pregnancy Diagnosis | 0.18 | 0.20 | 7.16% |
| Commercial Payer | 0.32 | 0.31 | -2.87% |
| Medicare Payer | 0.42 | 0.44 | 4.83% |
| Admissions Per Month | 8883 | 8348 | -6.03% |

**Note:** The second column is the average of the variable in the pre-disaster training data, while the third column is the average of the variable in the post-disaster test data. The fourth column is the percent difference from the pre-disaster training data to the post-disaster test data.

Table B-5: Changes in Case-Mix for Bellevue

| Variable | Training | Test | Percent Difference |
|---|---|---|---|
| Age | 53.83 | 55.10 | 2.35% |
| Age $< 18$ | 0.06 | 0.05 | -12.89% |
| Age $> 64$ | 0.38 | 0.41 | 9.03% |
| Diagnosis Acuity | 1.25 | 1.29 | 3.15% |
| Circulatory Diagnosis | 0.18 | 0.16 | -6.84% |
| Labor/Pregnancy Diagnosis | 0.17 | 0.19 | 10.79% |
| Commercial Payer | 0.24 | 0.24 | -2.08% |
| Medicare Payer | 0.39 | 0.42 | 9.23% |
| Admissions Per Month | 5140 | 4576 | -10.97% |

**Note:** The second column is the average of the variable in the pre-disaster training data, while the third column is the average of the variable in the post-disaster test data. The fourth column is the percent difference from the pre-disaster training data to the post-disaster test data.

Table B-6: Changes in Case-Mix for St. Johns

| Variable | Training | Test | Percent Difference |
|---|---|---|---|
| Age | 54.34 | 53.78 | -1.02% |
| Age $< 18$ | 0.05 | 0.05 | 11.83% |
| Age $> 64$ | 0.41 | 0.40 | -2.19% |
| Diagnosis Acuity | 1.23 | 1.27 | 3.14% |
| Circulatory Diagnosis | 0.17 | 0.18 | 5.38% |
| Labor/Pregnancy Diagnosis | 0.18 | 0.19 | 5.98% |
| Commercial Payer | 0.44 | 0.47 | 6.23% |
| Medicare Payer | 0.38 | 0.34 | -8.91% |
| Admissions Per Month | 3881 | 3626 | -6.58% |

**Note:** The second column is the average of the variable in the pre-disaster training data, while the third column is the average of the variable in the post-disaster test data. The fourth column is the percent difference from the pre-disaster training data to the post-disaster test data.

Table B-7: Changes in Case-Mix for Sumter

| Variable | Training | Test | Percent Difference |
|---|---|---|---|
| Age | 53.76 | 54.27 | 0.94% |
| Age $< 18$ | 0.07 | 0.04 | -44.86% |
| Age $> 64$ | 0.38 | 0.37 | -4.62% |
| Diagnosis Acuity | 1.24 | 1.29 | 3.71% |
| Circulatory Diagnosis | 0.16 | 0.18 | 11.41% |
| Labor/Pregnancy Diagnosis | 0.15 | 0.16 | 7.86% |
| Commercial Payer | 0.28 | 0.20 | -28.40% |
| Medicare Payer | 0.42 | 0.40 | -5.22% |
| Admissions Per Month | 496 | 424 | -14.40% |

**Note:** The second column is the average of the variable in the pre-disaster training data, while the third column is the average of the variable in the post-disaster test data. The fourth column is the percent difference from the pre-disaster training data to the post-disaster test data.

Table D-1: Predictive Accuracy using Percent Correct – Averaged over all Experiments

| Model | Percent Correct |
|---|---|
| Indic | 0.286 |
| | (0.280, 0.291) |
| Logit | 0.396 |
| | (0.391, 0.400) |
| Regular | 0.456 |
| | (0.451, 0.460) |
| DT | 0.441 |
| | (0.436, 0.445) |
| GBM | 0.464 |
| | (0.459, 0.469) |
| RF | 0.464 |
| | (0.459, 0.469) |
| Semipar | 0.414 |
| | (0.409, 0.419) |

**Note:** The table depicts average percent correct, averaged over all experiments. 95% confidence intervals computed from 500 bootstrap replications are in parentheses.

Table D-2: Predictive Accuracy using Percent Correct – Averaged over all Experiments – Relative to *Logit*

| Model | Relative Percent Correct |
|---|---|
| RRegular | 0.188 |
| | (0.174, 0.202) |
| DT | 0.149 |
| | (0.135, 0.163) |
| GBM | 0.205 |
| | (0.190, 0.219) |
| RF | 0.205 |
| | (0.191, 0.220) |
| Semipar | 0.057 |
| | (0.047, 0.067) |

**Note:** The table depicts the average percent correct, averaged across the different experiments, measured relative to the baseline parametric logit model *Logit*. 95% confidence intervals computed from 500 bootstrap replications are in parentheses.

Table D-3: Predictive Accuracy using Percent Correct – By Experiment

| Model | Sumter | StJohns | NYU | Moore | Coney | Bellevue |
|---|---|---|---|---|---|---|
| Indic | 0.462 | 0.200 | 0.197 | 0.196 | 0.331 | 0.326 |
| | (0.449, | (0.194, | (0.191, | (0.184, | (0.322, | (0.316, |
| | 0.475) | 0.206) | 0.203) | 0.209) | 0.341) | 0.336) |
| Logit | 0.615 | 0.319 | 0.404 | 0.295 | 0.336 | 0.405 |
| | (0.602, | (0.312, | (0.397, | (0.281, | (0.327, | (0.394, |
| | 0.628) | 0.325) | 0.412) | 0.309) | 0.346) | 0.415) |
| Regular | 0.589 | 0.375 | 0.458 | 0.480 | 0.363 | 0.468 |
| | (0.577, | (0.368, | (0.451, | (0.465, | (0.353, | (0.457, |
| | 0.602) | 0.382) | 0.465) | 0.495) | 0.373) | 0.478) |
| DT | 0.573 | 0.357 | 0.442 | 0.461 | 0.363 | 0.447 |
| | (0.559, | (0.350, | (0.435, | (0.446, | (0.354, | (0.437, |
| | 0.587) | 0.364) | 0.450) | 0.476) | 0.373) | 0.458) |
| GBM | 0.619 | 0.381 | 0.465 | 0.472 | 0.369 | 0.477 |
| | (0.607, | (0.374, | (0.458, | (0.456, | (0.359, | (0.467, |
| | 0.632) | 0.389) | 0.473) | 0.487) | 0.379) | 0.487) |
| RF | 0.628 | 0.380 | 0.459 | 0.481 | 0.370 | 0.467 |
| | (0.615, | (0.373, | (0.452, | (0.466, | (0.360, | (0.457, |
| | 0.641) | 0.387) | 0.467) | 0.496) | 0.379) | 0.477) |
| Semipar | 0.606 | 0.346 | 0.430 | 0.336 | 0.344 | 0.424 |
| | (0.593, | (0.339, | (0.422, | (0.322, | (0.334, | (0.414, |
| | 0.619) | 0.353) | 0.437) | 0.350) | 0.354) | 0.434) |

**Note:** The table depicts average percent correct by experiment. 95% confidence intervals computed from 500 bootstrap replications are in parentheses.

Table D-4: Predictive Accuracy using Percent Correct – By Experiment – Relative to *Logit*

| Model | Sumter | StJohns | NYU | Moore | Coney | Bellevue |
|---|---|---|---|---|---|---|
| Regular | -0.041 | 0.177 | 0.133 | 0.627 | 0.080 | 0.156 |
| | (-0.057, - | (0.155, | (0.116, | (0.559, | (0.057, | (0.132, |
| | 0.025) | 0.199) | 0.149) | 0.695) | 0.102) | 0.179) |
| DT | -0.068 | 0.120 | 0.093 | 0.563 | 0.081 | 0.105 |
| | (-0.086, - | (0.096, | (0.076, | (0.493, | (0.056, | (0.081, |
| | 0.050) | 0.143) | 0.111) | 0.633) | 0.106) | 0.129) |
| GBM | 0.008 | 0.197 | 0.150 | 0.598 | 0.098 | 0.178 |
| | (-0.009, | (0.174, | (0.132, | (0.526, | (0.075, | (0.154, |
| | 0.024) | 0.219) | 0.168) | 0.670) | 0.120) | 0.202) |
| RF | 0.022 | 0.192 | 0.136 | 0.629 | 0.100 | 0.154 |
| | (0.006, | (0.169, | (0.120, | (0.560, | (0.079, | (0.132, |
| | 0.037) | 0.214) | 0.152) | 0.699) | 0.121) | 0.176) |
| Semipar | -0.014 | 0.085 | 0.063 | 0.138 | 0.023 | 0.047 |
| | (-0.028, - | (0.063, | (0.050, | (0.095, | (0.003, | (0.031, |
| | 0.001) | 0.106) | 0.076) | 0.182) | 0.043) | 0.064) |

**Note:** The table depicts the average percent correct measured relative to the baseline parametric logit model *Logit* by experiment. 95% confidence intervals computed from 500 bootstrap replications are in parentheses.

Table D-5: Average Percent Improvement in Predictive Accuracy using Percent Correct, on the Training, Validation, and Test samples

| Model | Test | Validation | Train |
|---|---|---|---|
| Regular | 0.188 | 0.198 | 0.228 |
| | (0.174, 0.202) | (0.177, 0.219) | (0.217, 0.238) |
| DT | 0.149 | 0.211 | 0.325 |
| | (0.135, 0.163) | (0.190, 0.232) | (0.314, 0.337) |
| GBM | 0.205 | 0.249 | 0.366 |
| | (0.190, 0.219) | (0.227, 0.271) | (0.354, 0.378) |
| RF | 0.205 | 0.243 | 0.354 |
| | (0.191, 0.220) | (0.222, 0.264) | (0.343, 0.366) |
| Semipar | 0.057 | 0.088 | 0.131 |
| | (0.047, 0.067) | (0.075, 0.102) | (0.124, 0.138) |

**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. The training sample is a random 80% of the data pre-disaster, the validation sample a random 20% of the data pre-disaster, and the test sample data post-disaster. 95% confidence intervals computed from 500 bootstrap replications are in parentheses.

Table D-6: Average Percent Improvement in Predictive Accuracy for Percent Correct By Share of Destroyed Hospital

| Model | <10% | 10-30% | >30% |
|---|---|---|---|
| Regular | 0.233 | 0.163 | 0.128 |
| | (0.207, 0.260) | (0.145, 0.181) | (0.065, 0.191) |
| DT | 0.192 | 0.136 | 0.031 |
| | (0.166, 0.219) | (0.114, 0.157) | (-0.033, 0.095) |
| GBM | 0.241 | 0.202 | 0.117 |
| | (0.214, 0.268) | (0.182, 0.223) | (0.056, 0.177) |
| RF | 0.239 | 0.199 | 0.123 |
| | (0.213, 0.265) | (0.180, 0.218) | (0.065, 0.180) |
| Semipar | 0.082 | 0.064 | -0.075 |
| | (0.064, 0.100) | (0.050, 0.079) | (-0.123, -0.027) |

**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. Estimates are averaged across the experiments and broken down by the share of discharges of the destroyed hospital in the pre-disaster period predicted using the *Semipar* model. 95% confidence intervals computed from 500 bootstrap replications are in parentheses.

Table D-7: Percent Improvement in Predictive Accuracy for Percent Correct By Share of Destroyed Hospital, For Sumter Experiment

| Model | <15% | 15-50% | 50-80% | >80% |
|---|---|---|---|---|
| Regular | -0.052 | -0.024 | -0.064 | -0.026 |
| | (-0.081, -0.024) | (-0.079, 0.031) | (-0.095, -0.032) | (-0.042, -0.010) |
| DT | -0.051 | 0.025 | -0.157 | -0.074 |
| | (-0.086, -0.016) | (-0.030, 0.081) | (-0.199, -0.116) | (-0.096, -0.051) |
| GBM | 0.011 | 0.145 | -0.052 | -0.039 |
| | (-0.018, 0.040) | (0.087, 0.202) | (-0.084, -0.019) | (-0.058, -0.021) |
| RF | -0.001 | 0.132 | -0.008 | -0.008 |
| | (-0.030, 0.027) | (0.079, 0.186) | (-0.041, 0.024) | (-0.022, 0.005) |
| Semipar | -0.029 | 0.033 | -0.019 | -0.030 |
| | (-0.055, -0.002) | (-0.007, 0.074) | (-0.042, 0.004) | (-0.046, -0.014) |

**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. Estimates are for the Sumter experiment and broken down by the share of discharges of the destroyed hospital in the pre-disaster period predicted using the *Semipar* model. 95% confidence intervals computed from 500 bootstrap replications are in parentheses.

Table D-8: Predictive Accuracy using Percent Correct – Averaged over all Experiments – Relative to *Logit*, with Model Combination Model

| Model | Relative Percent Correct |
|---|---|
| Comb | 0.214 |
| | (0.199, 0.229) |
| Regular | 0.188 |
| | (0.174, 0.202) |
| DT | 0.149 |
| | (0.135, 0.163) |
| GBM | 0.205 |
| | (0.190, 0.219) |
| RF | 0.205 |
| | (0.191, 0.220) |
| Semipar | 0.057 |
| | (0.047, 0.067) |

**Note:** The table depicts the average percent correct, averaged across the different experiments, measured relative to the baseline parametric logit model *Logit*. 95% confidence intervals computed from 500 bootstrap replications are in parentheses. *Comb* is the Model Combination model using weights estimated on the 20% validation sample (allowing estimated weights to vary by disaster).

Table D-9: Predictive Accuracy using Percent Correct – By Experiment – Relative to *Logit*, with Model Combination Model

| Model | Sumter | StJohns | NYU | Moore | Coney | Bellevue |
|---|---|---|---|---|---|---|
| Comb | 0.009 | 0.205 | 0.153 | 0.627 | 0.109 | 0.182 |
| | (-0.007, | (0.182, | (0.135, | (0.555, | (0.088, | (0.158, |
| | 0.024) | 0.229) | 0.170) | 0.698) | 0.130) | 0.206) |
| Regular | -0.041 | 0.177 | 0.133 | 0.627 | 0.080 | 0.156 |
| | (-0.057, | (0.155, | (0.116, | (0.559, | (0.057, | (0.132, |
| | 0.025) | 0.199) | 0.149) | 0.695) | 0.102) | 0.179) |
| DT | -0.068 | 0.120 | 0.093 | 0.563 | 0.081 | 0.105 |
| | (-0.086, | (0.096, | (0.076, | (0.493, | (0.056, | (0.081, |
| | 0.050) | 0.143) | 0.111) | 0.633) | 0.106) | 0.129) |
| GBM | 0.008 | 0.197 | 0.150 | 0.598 | 0.098 | 0.178 |
| | (-0.009, | (0.174, | (0.132, | (0.526, | (0.075, | (0.154, |
| | 0.024) | 0.219) | 0.168) | 0.670) | 0.120) | 0.202) |
| RF | 0.022 | 0.192 | 0.136 | 0.629 | 0.100 | 0.154 |
| | (0.006, | (0.169, | (0.120, | (0.560, | (0.079, | (0.132, |
| | 0.037) | 0.214) | 0.152) | 0.699) | 0.121) | 0.176) |
| Semipar | -0.014 | 0.085 | 0.063 | 0.138 | 0.023 | 0.047 |
| | (-0.028, | (0.063, | (0.050, | (0.095, | (0.003, | (0.031, |
| | 0.001) | 0.106) | 0.076) | 0.182) | 0.043) | 0.064) |

**Note:** The table depicts the average percent correct measured relative to the baseline parametric logit model *Logit* by experiment. 95% confidence intervals computed from 500 bootstrap replications are in parentheses. *Comb* is the Model Combination model using weights estimated on the 20% validation sample (allowing estimated weights to vary by disaster).

Table D-10: Percent Improvement in Predictive Accuracy for Percent Correct for Previous Patients

| Model | All Patients | Previous Patients |
|---|---|---|
| Regular | 0.136 | 0.058 |
| | (0.125, 0.147) | (0.018, 0.098) |
| DT | 0.100 | -0.012 |
| | (0.088, 0.111) | (-0.054, 0.031) |
| GBM | 0.156 | 0.057 |
| | (0.144, 0.167) | (0.014, 0.100) |
| RF | 0.145 | 0.068 |
| | (0.134, 0.156) | (0.028, 0.108) |
| Semipar | 0.055 | -0.002 |
| | (0.045, 0.064) | (-0.039, 0.035) |

**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. Estimates are for the StJohns, Coney, NYU, and Bellevue experiments, and compare all patients to the identified set of patients that previously went to the destroyed hospital. 95% confidence intervals computed from 500 bootstrap replications are in parentheses.

Table D-11: Average Percent Improvement in Predictive Accuracy for Percent Correct By Share of Destroyed Hospital, by RF Minimum Node Size

| Model | <10% | 10-30% | >30% |
|---|---|---|---|
| RF | 0.239 | 0.199 | 0.123 |
| | (0.213, 0.265) | (0.180, 0.218) | (0.065, 0.180) |
| RF Small | 0.237 | 0.199 | 0.121 |
| | (0.211, 0.263) | (0.180, 0.218) | (0.064, 0.179) |
| RF Large | 0.171 | 0.144 | 0.118 |
| | (0.146, 0.196) | (0.126, 0.162) | (0.064, 0.172) |

**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. Estimates are averaged across the experiments and broken down by the share of discharges of the destroyed hospital in the pre-disaster period predicted using the *Semipar* model. *RF Large* model is the random forest model estimated using a minimum node size of 100, and the *RF Small* model is the random forest model estimated using a minimum node size of 10. 95% confidence intervals computed from 500 bootstrap replications are in parentheses.

Table D-12: Percent Improvement in Predictive Accuracy for Percent Correct By Share of Destroyed Hospital, For Sumter Experiment, by RF Minimum Node Size

| Model | <15% | 15-50% | 50-80% | >80% |
|---|---|---|---|---|
| RF | -0.001 | 0.132 | -0.008 | -0.008 |
| | (-0.030, 0.027) | (0.079, 0.186) | (-0.041, 0.024) | (-0.022, 0.005) |
| RF Small | -0.003 | 0.140 | -0.011 | -0.005 |
| | (-0.031, 0.026) | (0.087, 0.193) | (-0.043, 0.021) | (-0.018, 0.008) |
| RF Large | -0.096 | -0.037 | -0.109 | -0.048 |
| | (-0.129, -0.064) | (-0.088, 0.015) | (-0.144, -0.073) | (-0.067, -0.029) |

**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. Estimates are for the Sumter experiment and broken down by the share of discharges of the destroyed hospital in the pre-disaster period predicted using the *Semipar* model. *RF Large* model is the random forest model estimated using a minimum node size of 100, and the *RF Small* model is the random forest model estimated using a minimum node size of 10. 95% confidence intervals computed from 500 bootstrap replications are in parentheses.

Table D-13: Relative Improvement in Percent Correct – Damaged Areas

| Model | Sumter | StJohns | Coney |
|---|---|---|---|
| Regular | 0.014 | 0.078 | 0.179 |
| | (0.006, 0.021) | (0.046, 0.111) | (0.123, 0.236) |
| DT | -0.102 | 0.002 | 0.196 |
| | (-0.122, -0.081) | (-0.034, 0.037) | (0.133, 0.259) |
| GBM | -0.004 | 0.090 | 0.200 |
| | (-0.017, 0.010) | (0.057, 0.123) | (0.137, 0.264) |
| RF | -0.001 | 0.093 | 0.205 |
| | (-0.013, 0.010) | (0.062, 0.123) | (0.146, 0.265) |
| Semipar | 0.001 | -0.001 | 0.100 |
| | (-0.007, 0.010) | (-0.029, 0.027) | (0.041, 0.160) |

**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. Estimates are by experiment. Only zip codes with substantial disaster damage as indicated in Section C.1 are included. 95% confidence intervals computed from 500 bootstrap replications are in parentheses.

Table D-14: Relative Improvement in Percent Correct – Non-Damaged Areas

| Model | Sumter | StJohns | Coney |
|---|---|---|---|
| Regular | -0.098 | 0.225 | 0.039 |
| | (-0.128, -0.067) | (0.196, 0.254) | (0.018, 0.059) |
| DT | -0.033 | 0.178 | 0.033 |
| | (-0.065, -0.001) | (0.147, 0.209) | (0.008, 0.059) |
| GBM | 0.019 | 0.249 | 0.055 |
| | (-0.011, 0.050) | (0.219, 0.279) | (0.033, 0.077) |
| RF | 0.045 | 0.240 | 0.056 |
| | (0.015, 0.075) | (0.210, 0.270) | (0.037, 0.075) |
| Semipar | -0.030 | 0.127 | -0.009 |
| | (-0.056, -0.005) | (0.098, 0.155) | (-0.025, 0.007) |

**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. Estimates are by experiment. Only zip codes without substantial disaster damage as indicated in Section C.1 are included. 95% confidence intervals computed from 500 bootstrap replications are in parentheses.

Table D-15: Average Relative Improvement in Percent Correct: Emergency and MDC

| Model | Pregnancy | Cardiac | Non-Emer | Emer |
|---|---|---|---|---|
| Regular | 0.267 | 0.150 | 0.183 | 0.391 |
| | (0.234, 0.299) | (0.113, 0.188) | (0.163, 0.204) | (0.340, 0.441) |
| DT | 0.245 | 0.134 | 0.124 | 0.360 |
| | (0.211, 0.279) | (0.089, 0.180) | (0.101, 0.146) | (0.309, 0.411) |
| GBM | 0.257 | 0.179 | 0.209 | 0.402 |
| | (0.224, 0.289) | (0.136, 0.221) | (0.188, 0.230) | (0.350, 0.454) |
| RF | 0.282 | 0.162 | 0.207 | 0.397 |
| | (0.249, 0.314) | (0.121, 0.203) | (0.186, 0.227) | (0.346, 0.447) |
| Semipar | 0.064 | 0.024 | 0.052 | 0.110 |
| | (0.043, 0.085) | (-0.008, 0.055) | (0.039, 0.064) | (0.086, 0.135) |

**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. Estimates are averaged across experiments but separated by cardiac (MDC = 5), pregnancy (MDC = 14), emergency, and non-emergency patients. 95% confidence intervals computed from 500 bootstrap replications are in parentheses.

Table D-16: Average Relative Improvement in Percent Correct: Acuity

| Model | Low Acuity | Medium Acuity | High Acuity |
|---|---|---|---|
| Regular | 0.214 | 0.170 | 0.129 |
| | (0.196, 0.233) | (0.148, 0.193) | (0.092, 0.165) |
| DT | 0.174 | 0.128 | 0.096 |
| | (0.153, 0.194) | (0.105, 0.151) | (0.056, 0.135) |
| GBM | 0.224 | 0.191 | 0.163 |
| | (0.204, 0.244) | (0.168, 0.214) | (0.124, 0.202) |
| RF | 0.227 | 0.191 | 0.149 |
| | (0.208, 0.246) | (0.168, 0.214) | (0.113, 0.184) |
| Semipar | 0.073 | 0.037 | 0.052 |
| | (0.059, 0.086) | (0.023, 0.051) | (0.021, 0.082) |

**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. Estimates are averaged across experiments but separated by patients with different disease acuity (DRG weight) divided into terciles. 95% confidence intervals computed from 500 bootstrap replications are in parentheses.

Table D-17: Average Relative Improvement in Percent Correct: Payer Type

| Model | Commercial | Medicare | Medicaid | Medicare (Separate Est) |
|---|---|---|---|---|
| Regular | 0.208 | 0.160 | 0.317 | 0.117 |
| | (0.180, 0.235) | (0.141, 0.179) | (0.270, 0.364) | (0.098, 0.136) |
| DT | 0.181 | 0.121 | 0.284 | 0.139 |
| | (0.151, 0.212) | (0.102, 0.140) | (0.230, 0.338) | (0.118, 0.160) |
| GBM | 0.255 | 0.170 | 0.344 | 0.140 |
| | (0.225, 0.284) | (0.151, 0.190) | (0.291, 0.396) | (0.120, 0.160) |
| RF | 0.229 | 0.173 | 0.366 | 0.155 |
| | (0.202, 0.256) | (0.154, 0.191) | (0.316, 0.416) | (0.137, 0.174) |
| Semipar | 0.095 | 0.028 | 0.113 | 0.014 |
| | (0.074, 0.117) | (0.014, 0.041) | (0.080, 0.147) | (0.001, 0.027) |

**Note:** Percent correct measured relative to the baseline parametric logit model *Logit*. Estimates are averaged across experiments but separated by patients with different payers. For the Medicare (Separate Est) bars, we examine Medicare patients only and reestimate all of the models on the Medicare only sample in order to develop predictions. 95% confidence intervals computed from 500 bootstrap replications are in parentheses.

Table D-18: Predictive Accuracy using RMSE – Averaged over all Experiments – Relative to *Logit*

| Model | Relative RMSE |
|---|---|
| Regular | 0.0249 |
| | (0.0229, 0.0270) |
| DT | 0.0058 |
| | (0.0034, 0.0083) |
| GBM | 0.0367 |
| | (0.0346, 0.0387) |
| RF | 0.0373 |
| | (0.0357, 0.0389) |
| Semipar | 0.0020 |
| | (0.0005, 0.0035) |

**Note:** The table depicts the average RMSE, averaged across the different experiments, measured relative to the baseline parametric logit model *Logit*; since we report percent improvement, the negative is the change in RMSE. 95% confidence intervals computed from 500 bootstrap replications are in parentheses.

Table D-19: Predictive Accuracy using RMSE – By Experiment – Relative to *Logit*

| Model | Sumter | StJohns | NYU | Moore | Coney | Bellevue |
|---|---|---|---|---|---|---|
| Regular | -0.0666 | 0.0224 | 0.0397 | 0.0880 | 0.0209 | 0.0452 |
| | (-0.0748, -0.0585) | (0.0204, 0.0244) | (0.0370, 0.0423) | (0.0815, 0.0946) | (0.0187, 0.0230) | (0.0409, 0.0496) |
| DT | -0.1228 | 0.0048 | 0.0273 | 0.0795 | 0.0166 | 0.0296 |
| | (-0.1339, -0.1116) | (0.0023, 0.0074) | (0.0244, 0.0302) | (0.0724, 0.0865) | (0.0140, 0.0193) | (0.0250, 0.0341) |
| GBM | -0.0146 | 0.0244 | 0.0441 | 0.0911 | 0.0272 | 0.0477 |
| | (-0.0212, -0.0079) | (0.0223, 0.0266) | (0.0411, 0.0470) | (0.0836, 0.0986) | (0.0248, 0.0296) | (0.0431, 0.0522) |
| RF | -0.0072 | 0.0253 | 0.0408 | 0.0927 | 0.0251 | 0.0470 |
| | (-0.0132, -0.0013) | (0.0236, 0.0271) | (0.0385, 0.0431) | (0.0872, 0.0981) | (0.0231, 0.0272) | (0.0434, 0.0506) |
| Semipar | -0.0340 | 0.0066 | 0.0123 | 0.0141 | 0.0045 | 0.0086 |
| | (-0.0411, -0.0268) | (0.0049, 0.0084) | (0.0103, 0.0143) | (0.0103, 0.0178) | (0.0028, 0.0062) | (0.0061, 0.0111) |

**Note:** The table depicts the average RMSE measured relative to the baseline parametric logit model *Logit* by experiment; since we report percent improvement, the negative is the change in RMSE. 95% confidence intervals computed from 500 bootstrap replications are in parentheses.