WORKING PAPERS

Industrial Reorganization: Learning about Patient Substitution Patterns from Natural Experiments

Devesh Raval
Ted Rosenbaum
Nathan E. Wilson

WORKING PAPER NO. 329

May 2016

BUREAU OF ECONOMICS
FEDERAL TRADE COMMISSION
WASHINGTON, DC 20580

Industrial Reorganization: Learning about Patient Substitution Patterns from Natural Experiments*

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

May 25, 2016

Abstract

Despite their widespread usage, little is known about the predictive accuracy of different discrete choice demand models. To evaluate their performance, we use a series of natural disasters that unexpectedly removed hospitals from consumers' choice sets. We compare the model predictions of post-disaster behavior to the benchmark of actual post-disaster consumer behavior. Across our different settings, we find that models that allow for flexible interactions between patient characteristics and unobserved hospital quality perform the best and that it is important to use different classes of models. Further, the use of less accurate models could lead to more lax merger enforcement.

JEL Codes: C18, I11, L1, L41

Keywords: hospitals, natural experiment, patient choice, forecasting, antitrust

*We would like to thank Jonathan Byars, Gregory Dowd, Aaron Keller, Laura Kmitch, and Peter Nguon for their excellent research assistance. We also wish to express our appreciation for audiences and our discussants – Nathan Miller, Yair Taylor, Alex Fakos, and Kate Ho – at the 2015 AEA Meetings, 2015 DC IO Day, 2015 IIOC, and the 2015 FTC Microeconomics Conference. We would also like to thank Chris Garmon, Marty Gaynor, Dan Hosken, Ginger Jin, Francine Lafontaine, Jason O'Connor, Dave Schmidt, Bob Town, and anonymous FTC referees for their detailed comments on the paper. 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.

1 Introduction

Since the seminal work of McFadden (1981), discrete choice demand models have become one of the main tools used by economists, quantitative marketers, and statisticians to model consumer preferences. Underpinning the choice models' popularity and prominence is the comparative ease with which they permit counterfactual analysis of behavior. For example, economists use them to calculate the substitutability of merging parties' products in antitrust (Hendel and Nevo, 2006), to evaluate the extent of exchange rate passthrough in trade (Goldberg, 1995), and to assess how usage patterns would change following the introduction of a new light rail system in transportation policy (McFadden et al., 1977).

While they have proven popular in many fields, structural choice models' predictions rely heavily on functional form assumptions, the exogeneity of product characteristics, and the presence of a sufficient number of consumer and product characteristics to capture plausible substitution patterns. None of these criteria can typically be tested in the data, because quasi-experimental variation in choice sets is rare. For example, the entry and exit of products is usually endogenous. Indeed, Angrist and Pischke (2010) have suggested that the field of industrial organization's heavy reliance on unvalidated models is so problematic that it should be rebranded as *industrial disorganization*.

Even defenders of structural choice models acknowledge that "estimates driven by functional form rather than credible sources of identification in the data are unlikely to produce useful predictions" (Nevo and Whinston, 2010). However, in any given context, economists will debate what constitutes "credible sources of identification". Researchers may disagree on how much the standard logit functional form assumption drives results, what types of data are important to capture reasonable substitution patterns, and if there is sufficient variation in the instruments to identify any random coefficients.[1] In other words, it is difficult to

[1]The debate between Jerry Hausman and Tim Bresnahan on the welfare benefits of the introduction

ascertain which model is best and whether the best model is good enough. In most contexts, there is no way to settle these debates, since there is no objective benchmark with which to assess models.

In this paper, we use true natural experiments to evaluate models of consumer choice. In particular, we focus on consumer choice of hospitals before and after four natural disasters that severely damaged or destroyed hospitals but left the majority of the surrounding area undisturbed. These natural disasters exogenously altered consumers' choice sets, creating a benchmark against which to assess the performance of structural models. We use pre-disaster data to estimate consumer demand under different models used by researchers and policymakers and then predict the effects of the hospital elimination for each model. By comparing the models' predictions to actual post-disaster effects, we evaluate the relative strengths and weaknesses of various models. The natural experiments thus serve as a "laboratory" to assess model performance.

Across all our natural experiments, we confirm the importance of accounting for consumer heterogeneity in preferences for unobserved product quality.[2] In general, we find that the model that best predicts individual choices uses a semiparametric approach that allows for very flexible substitution patterns across consumers, and only imposes the logit independence of irrelevant alternatives (IIA) assumption across small and largely homogeneous groups. The other best performing model includes a large number of individual level demographics interacted with product dummy variables in order to account for this heterogeneity. In contrast, models that include only observable product characteristics perform worse than models that allow for preference heterogeneity across product fixed effects.

of Apple Cinnamon Cheerios is an example of such disagreement in a similar context (Hausman, 1997; Bresnahan, 1997).

[2] Ackerberg et al. (2007) state that "Attempts we have seen to model a random coefficient on the ξ [unobserved product quality] have lead to results which indicate that there was no need for one." This could suggest that it is unimportant to model heterogeneity in preferences for unobserved product quality. Our results suggest that, at least where there is rich micro data, it is *very* important to account for heterogeneous consumer preferences for unobserved quality.

However, even within the class of flexible models, we do not find a uniformly "preferred model." Rather, different models are better at predicting outcomes for different populations. For example, the semiparametric model overfits the data among groups where the popularity of the destroyed hospital was particularly high. Among such groups, more efficient parametric models produce more accurate results on average.

In the absence of a single "preferred model," we borrow an approach from macroeconomic forecasting to determine the optimal way to combine the predictions of the various models. Our results show that no single model strictly dominates the others, and so receives all of the weight, for any of the disasters. Rather, an optimal combination of models typically places positive weight on more than one approach, with roughly equal weight on the predictions from a semiparametric model and flexible parametric models. These results suggest that both parametric and semiparametric approaches, or a model combination of them, be used in practice to ensure that policy decisions are robust to either the overfitting of a semiparametric model or the inflexibility of a given parametric model.

Within the hospital context, the heterogeneity in the settings we study gives us confidence that our findings are relevant beyond these settings. The natural disasters we study occurred in rural, suburban, and urban environments, and the destroyed hospitals range from NYU Langone – a large, nationally ranked teaching hospital – to small community hospitals. Beyond the health care context, our results provide guidance to researchers applying discrete choice models to other areas where rich micro data are available, such as e-commerce (e.g., Einav and Levin, 2014), durable goods (e.g., Berry et al., 2004), telecommunications (e.g., Goolsbee and Petrin, 2004), neighborhood choice (e.g., Bayer et al., 2007), and political campaigns (e.g., Gordon and Hartmann, 2013). While we compare different model specifications for hospital choice, researchers studying other settings are likely to face similar decisions in accounting for unobserved product quality and in the degree of structure to impose on the data.

In this way, we contribute to the small but growing literature that examines the performance of structural econometric models using experimental variation. LaLonde (1986) is one of the first papers to do so, comparing structural models of the effects of a job trainee program to the results of a field experiment. Other papers specifically examining discrete choice models are Todd and Wolpin (2006), who estimate a model of child schooling and fertility using pre-treatment data and then compare its predictions to the behavior of an experimental group, Conlon and Mortimer (2013), who compare the predictions of parametric choice models to the behavior observed when products are removed from vending machines, and Pathak and Shi (2014), who estimate a model predicting school choices in Boston and plan to evaluate forecasts from that model after a major policy change.

We also aim to provide guidance to researchers studying competition in health care markets (e.g., Garmon, 2016; Ho and Lee, 2015; Shepard, 2016). The extent of a managed care organization's bargaining leverage vis a vis a hospital in negotiating their contract depends upon how consumers react if a hospital is removed from an insurance plan (Capps et al., 2003; Gowrisankaran et al., 2015). Our natural experiments examine how well different models capture patients' substitution patterns after a hospital was removed from their choice set, and thus directly address this question.

Finally, we examine whether model choice matters for the policy counterfactual of the welfare impact of hospital mergers. Specifically, we consider the connection between model accuracy and the welfare effects of a series of simulated hospital mergers. Dividing our simulated mergers into two groups based on the magnitude of predicted harm, we find that the standard deviation of the percent change in the loss of consumer welfare across models is approximately 15% of the average for mergers that have larger predicted consumer harm.[3]

[3]We use the willingness to pay approach from Capps et al. (2003), as explained in Gowrisankaran et al. (2015). We separate the mergers into two groups with greater and smaller WTP based on a WTP threshold of 5%. There are alternative approaches to measure changes in consumer welfare and hospital prices from mergers as in Ho and Lee (2015) and Dafny et al. (2016).

Such variation is of significant magnitude to provide qualitatively different predictions of competitive effects. Further, we find that, on average, failing to use a better fitting model leads to underpredictions of consumer harm for those mergers most likely to harm consumers. Therefore, in our sample, errors made as the result of using worse fitting models could underestimate the reduction in competition from mergers.

The paper proceeds as follows. Section 2 describes our experimental design. In Section 3, we briefly lay out the theoretical framework underpinning the modern approach to estimating patient choice models, and discuss the specifications we focus on in this paper. In Section 4, we present our results on model performance. We examine policy counterfactuals from mergers in Section 5. Section 6 provides an overview of a set of robustness checks that we have done, and Section 7 concludes.

2 Natural Experiments

2.1 Disasters

For our natural experiments, we exploit the unexpected closures of six hospitals in four different markets following a natural disaster. Our set of disasters is itemized below in Table I. The Americus tornado struck a community hospital in rural Georgia, while the Moore tornado hit the suburbs of Oklahoma City. Hurricane Sandy flooded portions of Manhattan and Brooklyn in New York City, and closed three hospitals in those boroughs. The Northridge earthquake hit Los Angeles and closed one hospital in the Santa Monica area.

For a natural disaster to provide a good natural experiment to assess choice models, it must satisfy several criteria. First, the service area must be large enough and the period post disaster for which the hospital is closed long enough that we have enough power to compare

Table I 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

different demand models. Second, the destroyed hospital must have had a large enough market share in the service area for the disaster, because the experiment is informative on model performance only when the choice environment undergoes a substantial change. Finally, the damage from the disaster must be narrow enough that the change in patient decision making is limited to the change in the choice set.

In addition, experiments should have greater external validity to other settings if there is greater heterogeneity in the treated groups. In our case, the more the characteristics of the destroyed hospital and patient choice sets vary considerably across disasters, the more we are comfortable generalizing our results beyond our specific settings.

In the next sections, we demonstrate both that our disasters are good natural experiments, and discuss the heterogeneity that makes us more comfortable extrapolating our results to other settings.

2.2 Service Areas

We first construct a service area for each hospital in order to assess whether our criteria are met. To construct the service area for each destroyed hospital, we determine the set of patients that were likely to consider the destroyed hospital by looking at all patients going to general acute care hospitals living in zip codes that comprise a 90% service area of the

destroyed hospital.[4] To compute this service area, we rank all zip codes by the number of patient discharges from the destroyed hospital. Then, beginning with the zip code with the most patients, we add zip codes until the sum of the hospital's patients from those zip codes is above 90% of overall discharges. This approach may sweep in many zip codes where the hospital is competitively insignificant. Therefore, we exclude any zip code where the hospital's share in the pre-disaster period is below 4%.[5]

Table II displays some characteristics from the service area of each destroyed hospital that allow us to assess whether the disaster provide good natural experiments. The service area for Sumter Regional experiences a massive change post-disaster, as the share of the destroyed hospital is over 50 percent. For the other disasters, the share of the destroyed hospital roughly ranges from 9 to 18 percent. Thus, the destroyed hospital has a large enough share in each service area that patients' choice environment changes substantially.

To construct the post-disaster period for analysis, we exclude the period immediately surrounding the disaster from our analysis.[6] We do so in order to prevent both injuries from the disaster and changes in traffic patterns following the disaster due to residual damage from affecting our analysis. We have a substantial number of patient admissions post-period after each disaster in which to examine model performance, ranging from about four to five thousand admissions for Moore and Sumter, nine to ten thousand for Bellevue and Coney, and fifteen to twenty thousand for NYU and St. John's.[7] Thus, we have enough admissions in the post-period for all of the disasters to have power to compare different discrete choice models.

[4]Our primary source of data for each of our natural experiments is the inpatient hospital discharge data collected by state departments of health. Such patient-hospital data have been previously used by researchers (Capps et al., 2003; Ciliberto and Dranove, 2006), and provide a host of characteristics describing the patient receiving care as well as the type of clinical care being provided. The details on the construction of our estimation samples are provided in Appendix B.

[5]Our results are robust to changes in this cutoff.

[6]We describe the periods dropped for each disaster in Appendix B.

[7]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 about a 3 percent share in the Coney service area.

Table II Descriptive Statistics of Affected Hospital Service Areas

Event	Service Area Share Destroyed Hospital	Post-Period Admissions	2010 Pop	Hospitals	Beds Destroyed	Average Beds Other
Sumter	50.4%	5,092	56,485	15	132	252
NYU	8.9 %	16,696	1,356,836	19	791	744
Coney	18.2%	9,666	613,894	17	371	831
Bellevue	10.8%	9,152	785,462	20	807	723
Moore	11.0%	4,480	98,976	12	45	350
StJohns	17.4%	18,130	686,031	21	610	401

Note: The first column indicates the share of the destroyed hospital in the service area, the second column the total number of admissions in the post-period, and the third column the 2010 Census population of the service area. The fourth column is the number of hospitals in the service area, the fifth column is the number of beds in the destroyed hospital, and the sixth column is the average number of beds in all other hospitals. Sources: State hospital discharge data, AHA, Census.

2.3 Choice Sets

Then, for that set of patients, we determine the choice set as the hospitals that are likely to be the destroyed hospitals' closest substitutes. We define the set of relevant substitute hospitals by including all hospitals in the choice set that have a share of above 1% for the patients in the 90% service area as defined above in a given month (quarter for the smaller Sumter and Moore). Any hospital that does not meet this threshold is included in the "outside option." We examine the robustness of our results to choice set changes in Appendix D.7.

For external validity, it is useful to have our disasters affect different types of hospitals and different settings. Table II provides insight into the large variation in market characteristics across our different settings. Using population estimates from the 2010 Census, the rural area of Sumter has about 55,000 people and the suburbs of Moore 100,000. In the densely populated urban core of New York City, the Coney service area has over 600,000 people, the Bellevue service area over 750,000, and NYU 1.35 million. The St. John's service area in Southern California has about 700,000 people in 2010.

For Sumter, the destroyed hospital is a small community hospital with 132 beds; the surrounding hospitals are on average double its size, but there are no high quality teaching

hospitals. For Moore, the destroyed hospital is extremely small at 45 beds, and nearby hospitals are on average almost ten times larger. One teaching hospital, OU Medical Center, is in the service area. The New York destroyed hospitals are all large – NYU and Bellevue are both over 700 beds each, and Coney is almost 400 beds – with several teaching hospitals in the service area, including NYU. St. John's is similarly large with 610 beds, and with several large and high quality hospitals in the area, including UCLA Medical Center. The heterogeneity across these environments gives us more confidence to extrapolate any consistent findings beyond these particular settings.

2.4 Disaster Damage

An important question for the validity of our natural experiments is whether or not the disaster meaningfully disrupted the wider area as well as causing the closure of particular facilities. In such a circumstance, we might reasonably be concerned that predictions based on the pre-period would not be meaningful following the disaster. Therefore, we have carefully assessed whether or not the disasters' impact could reasonably be characterized as narrow. Below, we present graphical evidence of the comparatively modest scope of damage in Sumter, Moore, Manhattan (where NYU Langone and Bellevue Hospital Center are located), Coney Island, and Los Angeles in Figure 1 - Figure 5. In each figure, the service area is shaded with zip codes with more overall hospital admissions in darker shading.

In all four circumstances, the figures suggest that the extent of the damage was comparatively limited compared to the size of the affected hospitals' service areas. For example, Figure 1 shows the path of the storm that destroyed Sumter Regional Hospital as a green line. Its path was very narrow, cutting through Americus city without affecting the rural areas surrounding Americus. The Moore tornado had a similar effect for the city of Moore relative to its suburbs.

The flooding from Hurricane Sandy – depicted in blue cross-hatching – primarily affected areas adjacent to water. The actual damage in Manhattan was fairly small, while in 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. Our review of damage data ultimately led us to conclude that we should drop Long Beach Medical Center, a fourth hospital closed due to Sandy, from our sample of disrupted hospitals. We found that a considerable portion of the Long Beach area experienced massive flood damage, and news reports indicate that life on Long Island was substantially disturbed for a protracted period following the storm.

Figure 5 shows that the damage in the Los Angeles area was more widespread than the other disasters; we depict the intensity of earthquake shaking in cross-hatching. While the Santa Monica area was particularly hard hit, many areas nearby received little structural damage from the earthquake.

As a robustness exercise, we have also explored dropping those zip codes that our research indicates were most severely impacted, which is presented in Appendix D.2.[8]

3 Model

3.1 Patient Choice

We want to understand how well econometric models capture a patient's decision of where to receive care in the event that their preferred hospital is eliminated from the choice set. Our representation of the patient choice process follows the prior literature, including Capps et al. (2003), Ho (2006), and Gowrisankaran et al. (2015).

A patient i becomes ill with condition c at time t in market m, and chooses the specific hospital h from the set of hospitals H $(h = 1, ..., N)$ that are available to them based on the

[8]To determine these areas, we draw on resources distributed by various federal, state, and local authorities acting in response to the disasters, such as damage maps, as well as surrounding media coverage.

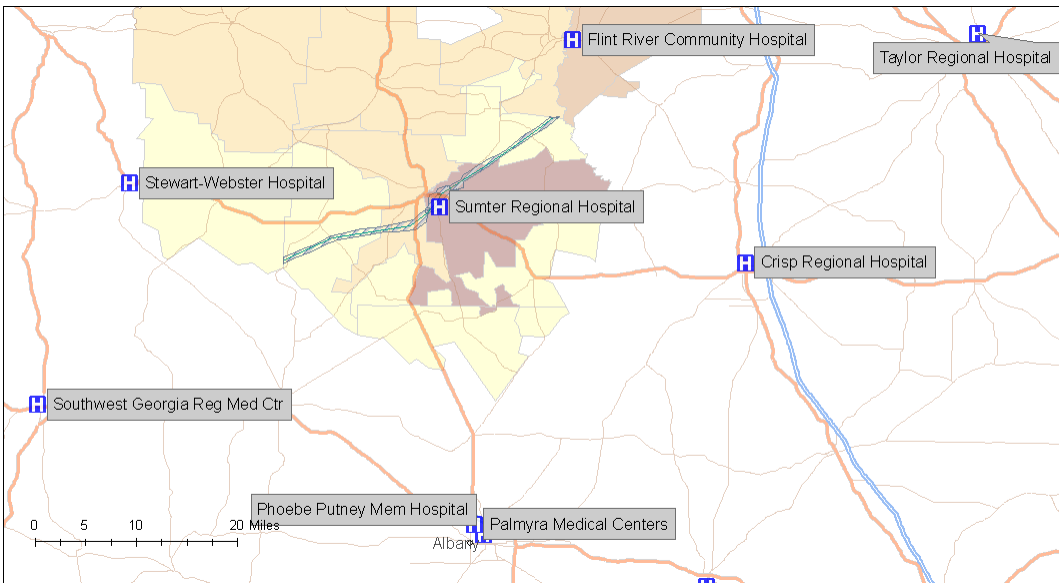


Figure 1 Damage Map in Americus, GA

Note: Green line indicates the path of the tornado and the shaded area around it is the government designated damage area. Darker shaded polygons are zip codes we use in the estimation. Zip codes with more overall hospital admissions during the estimation period are shaded in a darker color. Sources: City of Americus, GA Discharge Data

level of utility that they anticipate from receiving care there. We follow the existing literature in assuming that the patient's utility is a linearly separable combination of observable elements and an idiosyncratic shock. Specifically, this means that the utility patient i with condition c receives from care at hospital h can be represented as:

$$u_{ihc} = \delta_{ihc} + \epsilon_{ihc}, \tag{1}$$

where δ_{ihc} is the observable component of the patient's utility and ϵ_{ihc} is an unobserved shock affecting the relative likelihood that patient i chooses hospital h.

The existing literature on hospital choice typically assumes that after using "sufficient" observable variables x_{ihc} to generate δ_{ihc}, ϵ can be assumed to be independent and identically distributed random variables drawn from the type-I extreme value distribution. This implies that consumers with identical δs will on average exhibit similar preference patterns which

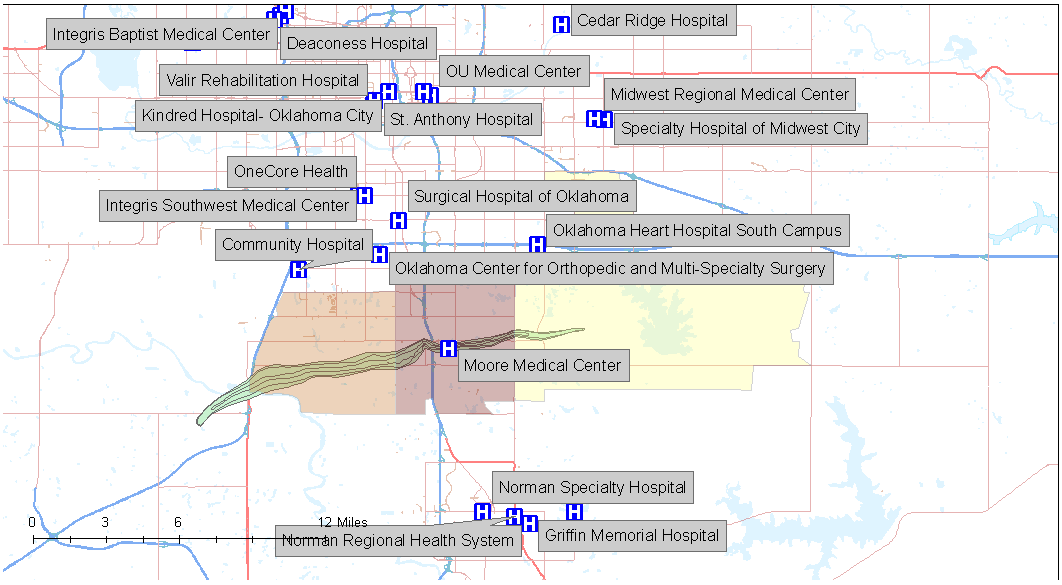


Figure 2 Damage Map in Moore, OK

Note: Green area indicates the damage path of the tornado. Darker shaded polygons are zip codes we use in the estimation. Zip codes with more overall hospital admissions during the estimation period are shaded in a darker color. Sources: NOAA, OK Discharge Data

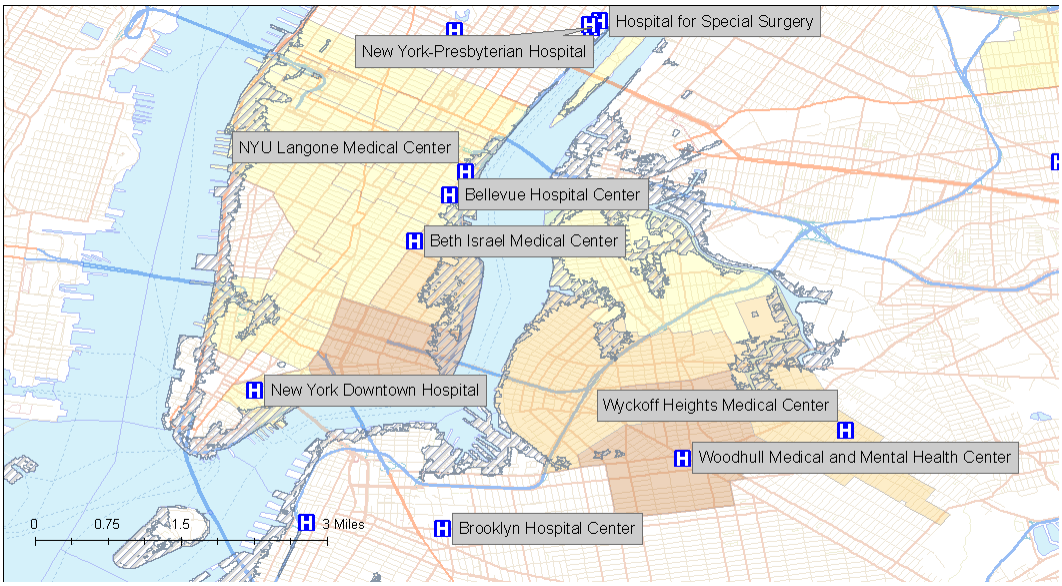


Figure 3 Damage Map in Manhattan, NY

Note: Cross-hatching indicates flood-affected areas. Darker shaded polygons are zip codes we use in the estimation for Bellevue. Zip codes with more overall hospital admissions during the estimation period are shaded in a darker color. Sources: FEMA, NY Discharge Data

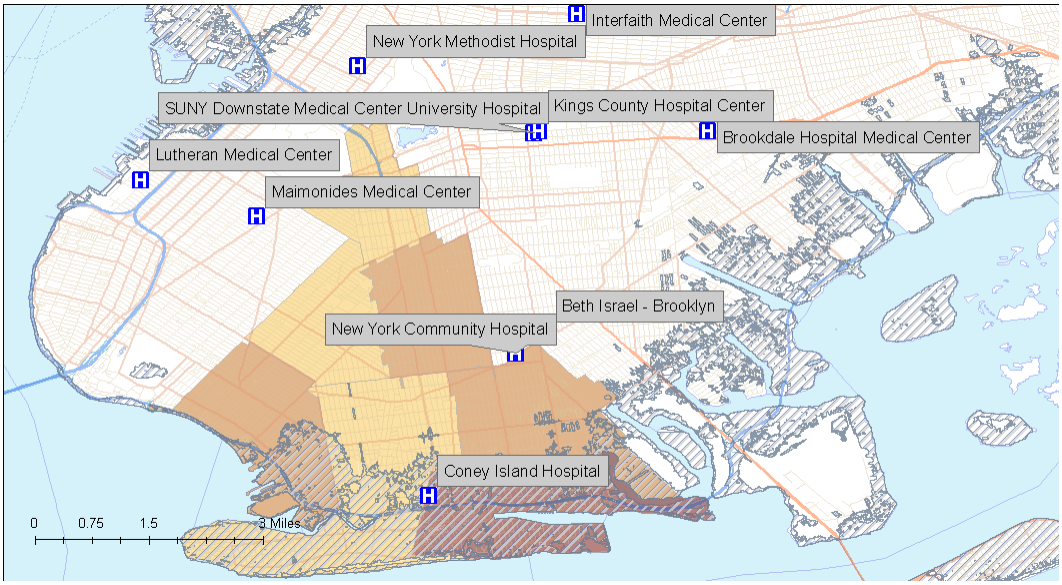


Figure 4 Damage Map in Coney Island, NY

Note: Cross-hatching indicates flood-affected areas. Darker shaded polygons are zip codes we use in the estimation. Zip codes with more overall hospital admissions during the estimation period are shaded in a darker color. Sources: FEMA, NY Discharge Data

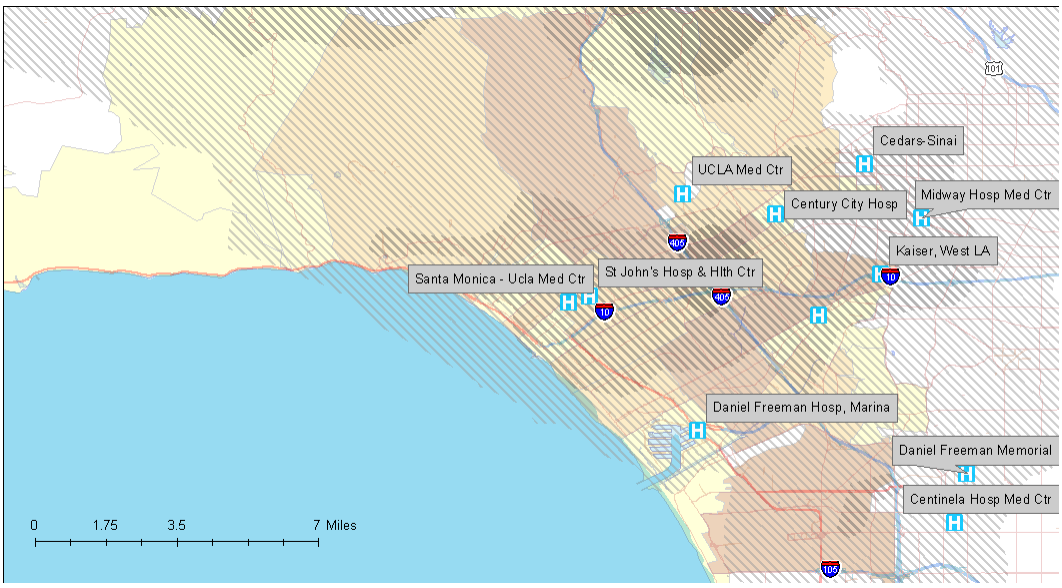


Figure 5 Damage Map in Los Angeles, CA

Note: Cross-hatching indicates 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 cross-hatched 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 cross-hatched. Darker shaded polygons are zip codes we use in the estimation. Zip codes with more overall hospital admissions during the estimation period are shaded in a darker color. Sources: USGS Shakemap, OSHPD Discharge Data

are independent of irrelevant alternatives (IIA). Moreover, the logit assumption implies that the probability that patient i with condition c receives care at hospital h is:

$$s_{ihc} = \frac{\exp(\delta_{ihc})}{\sum_{j \in H} \exp(\delta_{ijc})}. \tag{2}$$

3.2 Empirical Implementation

Researchers have placed very different assumptions on the most effective way to model δ in equation (1). Table III details some of the most prominent differences between the models that we examine by showing how they differ in their treatment of three important sets of variables: travel time, hospital characteristics, and hospital indicators.[9] One check mark

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

indicates the presence of an element. More check marks indicate the degree of interactions between that element and patient characteristics, which may include race, sex, and age, as well as the different diagnoses and procedures they have and their relative severity. Appendix C contains a detailed discussion of each model and Appendix F details all of the variables present in each model.

Table III Summary of Tested Models

Name	Travel Time	Hospital Characteristics	Hospital Indicators
Indic	✗	✗	✓
Char (**Garmon WP**)	✓✓	✓✓	✗
CDS (**RAND '03**)	✓✓	✓✓✓	✗
Time (**May WP**)	✓	✗	✓
Ho (**JAE '06**)	✓	✓✓✓	✓
GNT (**AER '15**)	✓✓	✓✓	✓✓
Inter	✓✓✓	✗	✓✓✓
Semipar (**Raval, Rosenbaum, Tenn WP**)	Hospital Indicators Interacted with Bins		

Note: Each row is a stylized depiction of a given model. More check marks indicate the degree of interactions between that element and patient characteristics, which may include race, sex, and age, as well as the different diagnoses and procedures they have and their relative severity.

Our reference model (*Indic*) assumes that there is no patient level heterogeneity. In other words, everyone within the relevant area has, on average, the same preferences for each hospital. As a result, patient choices can be modeled as being proportional to aggregate market shares, and δ can be estimated using only hospital indicators as covariates. In other words, this model could be estimated with only aggregate data.

In our view, the most important differentiator among the different models accounting for patient-level heterogeneity is whether or not they assume that patients' choices can be modeled exclusively in "characteristic" space (Lancaster, 1966; Aguirregabiria, 2011). That

is, unobserved hospital characteristics do not consistently affect patient preferences. We include two models (*CDS* and *Char*) that make this strong assumption, modeling δ simply as a function of a rich set of interaction terms between patient attributes (age, sex, income, condition, diagnosis) and hospital characteristics (for-profit status, teaching hospital, nursing intensity, presence of delivery room, etc.), as well as a measure of the travel time from the patient's home to the hospital.[10] *CDS* is based on Capps et al. (2003), while *Char* is based on Garmon (2016).

A contrasting set of specifications relaxes the strong assumption of no unobserved hospital characteristics made by the characteristic space models by including a hospital dummy, possibly interacted with individual-level characteristics. *Time* just includes hospital indicators, travel time, and travel time squared; May (2013) claims that this model performs just as well as more complicated models. *Ho*, based on Ho (2006), includes hospital indicators, as well as many interactions between hospital characteristics and patient characteristics in a similar way as the characteristics models. *GNT*, based on Gowrisankaran et al. (2015), includes a large set of interactions between travel time and patient characteristics, as well as hospital indicators and hospital characteristics interacted with acuity. *Inter* includes interactions of hospital indicators with acuity, major diagnostic category, and time as well as many interactions between patient characteristics and travel time.

Finally, we use a semiparametric bin estimator (*Semipar*), similar to Raval et al. (2015), which moves away from a characteristics based approach altogether to even more flexibly account for consumer heterogeneity across choices. This approach creates small and homoge-

[10] We obtain these hospital characteristics from the annual hospital characteristics data provided by the American Hospital Association (AHA) Guide and Medicare Cost Reports; they include such details as for-profit status, whether or not a hospital is an academic medical center or a children's hospital, the number of beds, the ratio of nurses to beds, the presence of different hospital services such as an MRI or cardiac ICU, and the number of residents per bed. For a few hospitals in California, New York and Oklahoma, the AHA and Medicare Cost Reports only contain data on the total hospital system rather than individual hospitals. For the AHA Guide, see http://www.ahadataviewer.com/book-cd-products/AHA-Guide/. For the Medicare Cost Reports, see http://www.cms.gov/Research-Statistics-Data-and-Systems/Files-for-Order/CostReports/index.html?redirect=/costreports/.

neous groups based upon patient characteristics, including zip code, age, disease acuity, and diagnosis category. The main assumption is that IIA holds within groups, and so hospital choice probabilities change proportionally to the observed shares of the group with a change in the choice set. In our implementation of this approach, we allow for group sizes as small as twenty, such that for some groups very few patients are used to predict substitution patterns. As discussed in Carlson et al. (2013) and Raval et al. (2015), this flexible approach is computationally efficient despite being equivalent to including a fixed effect for each group-hospital interaction in a multinomial logit model and allowing thousands of groups. While this model removes the restrictions of a characteristics based approach (for both products and consumers), it also has the potential to give extremely noisy estimates due to the very small group sizes.

All of these models provide different tradeoffs between model flexibility and power. A more flexible model is better able to account for consumer heterogeneity, and so has less bias, but its estimates are also likely to have greater variance. Since both greater bias and greater variance reduce out of sample performance, it is not clear which models will perform best.

4 Prediction

We estimate all of the models in Section 3 on data from the period before the disaster, and assess each model's predictive performance on data from the period after the disaster. Each model is thus assessed out of sample along two dimensions; first, it is estimated on an earlier time period, and second, the choice set available to patients has changed with the disaster. The change in the choice set is crucial to see how well each model predicts patients' choices after a major change in market structure.

4.1 Relative Performance

We compare the relative performance of the models on their predictions of aggregate market shares, aggregate diversion ratios post-disaster, and individual hospital choices for each destroyed hospital's service area.

4.1.1 Aggregate Shares

A simple way to assess performance on aggregate shares is to plot the time series of predictions against observed shares. In Figure 6, we do this for the Sumter disaster for three models, *Semipar*, *Inter*, and *CDS*, and 6 hospitals; the observed shares are the dashed red line. The grey dot-dash vertical line depicts the quarter of the disaster.

With the disaster, Sumter Regional's market share falls from about 50 percent to zero. Both the *Semipar* and *Inter* models closely track the actual changes in market shares for most of the remaining hospitals; for example, both predict the large rise in share for Phoebe Putney. *CDS*, on the other hand, performs very poorly; for example, it underpredicts the post-disaster share for Phoebe Putney by about 20 percentage points and predicts an initial share that is much higher than the observed share for Crisp Regional. All three models overpredict the share going to the outside option.

We examine the performance of all of the models across all of the destroyed hospitals using the criterion of root mean squared error (RMSE). At the aggregate level, the RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{N_J}\sum_j [y_j - \hat{y}_j]^2}.$$

Here y_j is the share of alternative j, $\hat{y}_j$ the model prediction, and N_J the total number of alternatives.

To look at relative differences across models, we examine the percent improvement in RMSE for each model over the baseline of the *Indic* model. The *Indic* model provides a

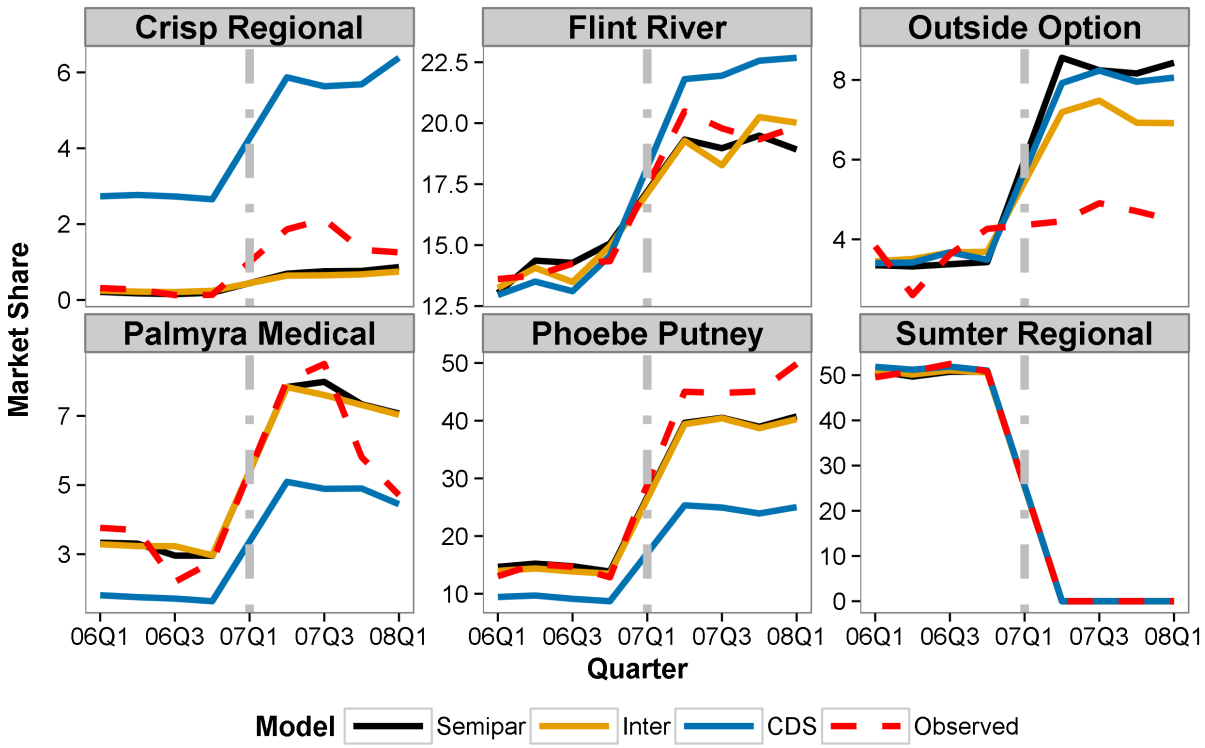
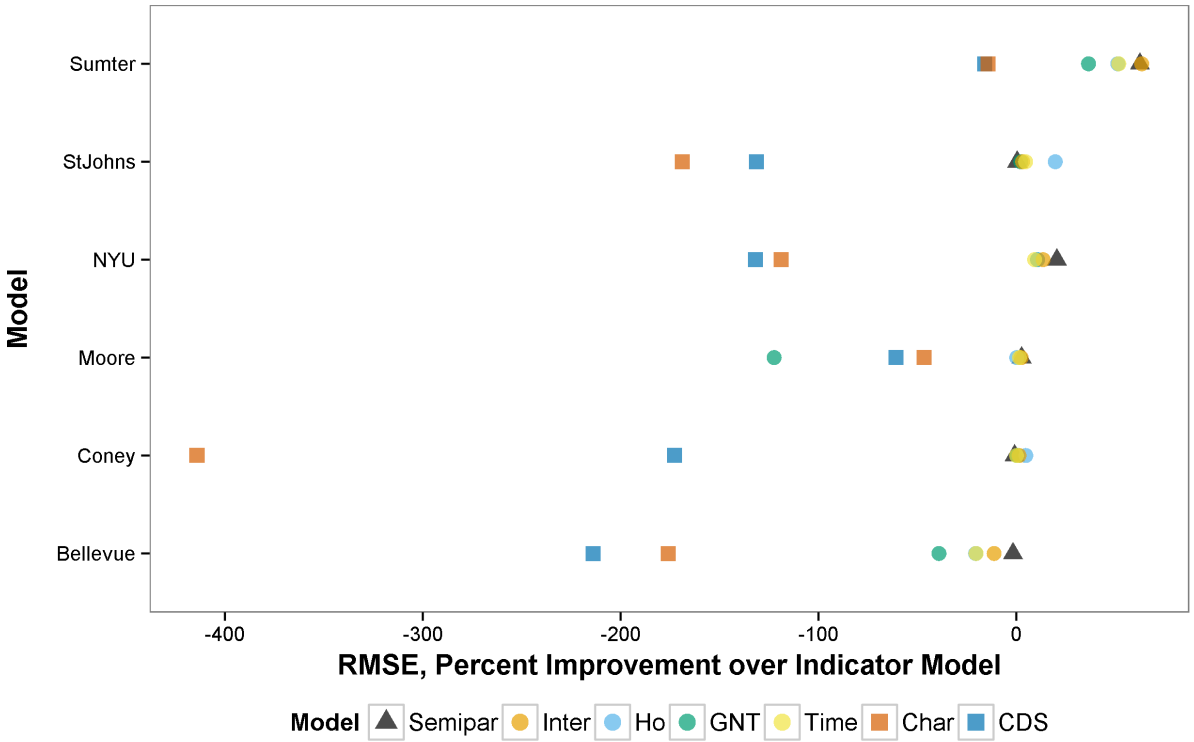
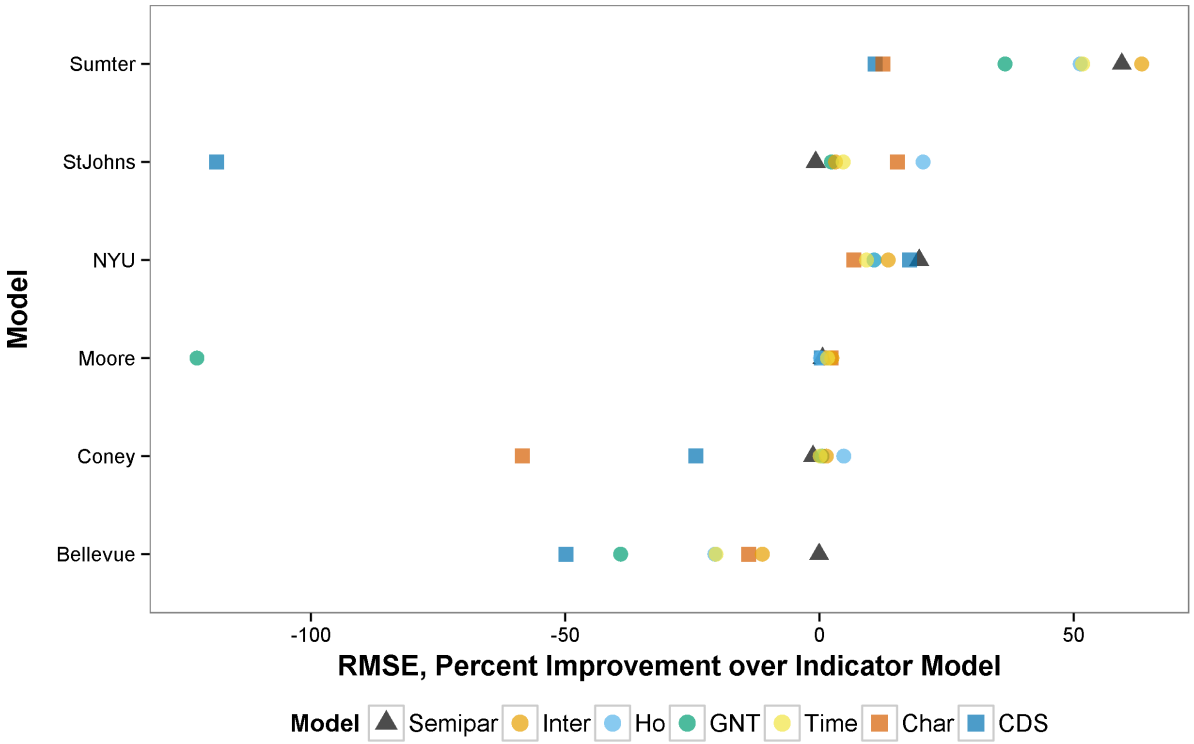


Figure 6 Aggregate Market Shares, Predicted and Observed, for Sumter

Note: Red dashed line is the observed series of market shares. The grey vertical dot-dash line depicts the quarter of the disaster.



(a) Aggregate Share



(b) Aggregate Diversion Ratio

Figure 7 Relative Improvement in RMSE of Aggregate Predictions

Note: Improvement is Percentage Improvement in RMSE for each model over the *Indic* model. Parametric models without hospital indicators are squares, with hospital indicators are circles, and semiparametric models are triangles.

useful baseline as it is a simple model that only requires aggregate data on market shares. We define the percent improvement as:

$$1 - \frac{RMSE_{Model}}{RMSE_{Indic}}.$$

Our results are shown in Figure 7a, which depicts the relative improvement in RMSE for each destroyed hospital's service area and model in the period after the disaster. Each row is a different destroyed hospital's service area; the models are distinguished both by color, and by shape, with the two parametric models without hospital indicators as squares, the four parameteric models with hospital indicators as circles, and the semiparametric model as a triangle.

The two characteristics based models – *CDS* and *Char* – always have much higher RMSE than the other models. Their RMSE is substantially worse than *Indic* for all of the service areas, and more than 100 percent worse for four service areas. These results demonstrate that unobserved hospital characteristics are important to understand aggregate hospital demand.

Across all of the hospitals, the differences between models that include hospital product effects – *Semipar*, *Inter*, *Ho*, *GNT*, and *Time* – are much smaller. For example, *Semipar* ranges from 20 percentage points worse than *Ho* to 20 percentage points better across the hospitals. Of these models, *Semipar*, *Inter*, and *Ho* are each the best model in two cases, although all models underperform our baseline model, *Indic*, for Bellevue.

These results illustrate the extent to which models match the levels of consumer choice probabilities before and after a choice was eliminated. However, in many applications, the *change* in consumers' choice probabilities for different options after removing an object from the choice set is the object of interest, such as the diversion ratio referred to in Garmon (2016). Therefore, we examine the RMSE of the aggregate diversion ratio following the

disaster. We define the aggregate diversion ratio for hospital j as:

$$\frac{y_{j,1} - y_{j,0}}{y_{dest,0}}$$

where $y_{j,1}$ is the share of hospital j in the period after the disaster, $y_{j,0}$ the share of hospital j before the disaster, and $y_{dest,0}$ the share of the destroyed hospital. Assuming that all changes in market shares after the disaster are due to the closure of the destroyed hospital, the diversion ratio tells us the fraction of the destroyed hospital's patients that went to hospital j. For the New York hospitals, the denominator of the diversion ratio includes all destroyed hospitals in the choice set.

Figure 7b depicts the relative improvement in RMSE for each model over *Indic*. We can immediately see that the characteristics models do much better on diversion ratios than they did on shares; their earlier poor performance stems primarily from missing the levels of shares. However, they are still typically worse than *Semipar* and *Inter*. Again, *Semipar*, *Inter*, and *Ho* are each the best model in two cases, although *Semipar* is second to *Indic* for Bellevue.

4.1.2 Individual Predictions

Since the shape of demand is determined by individual heterogeneous consumers, predictions on individual choice are key for assessing welfare in differentiated product markets. Figure 8 depicts the percent improvement over *Indic* for all models and across all of the hospitals for individual choices. We again measure model performance using RMSE, although we find in Appendix D.8 general agreement across alternative performance metrics.[11] The models that provide the greatest flexibility to capture unobserved patient heterogeneity always perform the best, on average, across all of our experimental settings: *Semipar* and *Inter* are always

[11] These alternative metrics are Mean Absolute Error, zero-one loss based on whether the patient went to the choice with the highest probability, and relative entropy (a log likelihood based statistic).

the best and second-best performing models. *Inter* is the best performing model for Sumter, while *Semipar* performs best for the other five destroyed hospitals. Across hospitals, *Semipar* is 0.5 to 3 percentage points better than *Ho*, and between one to 13 percentage points better than *Indic*.

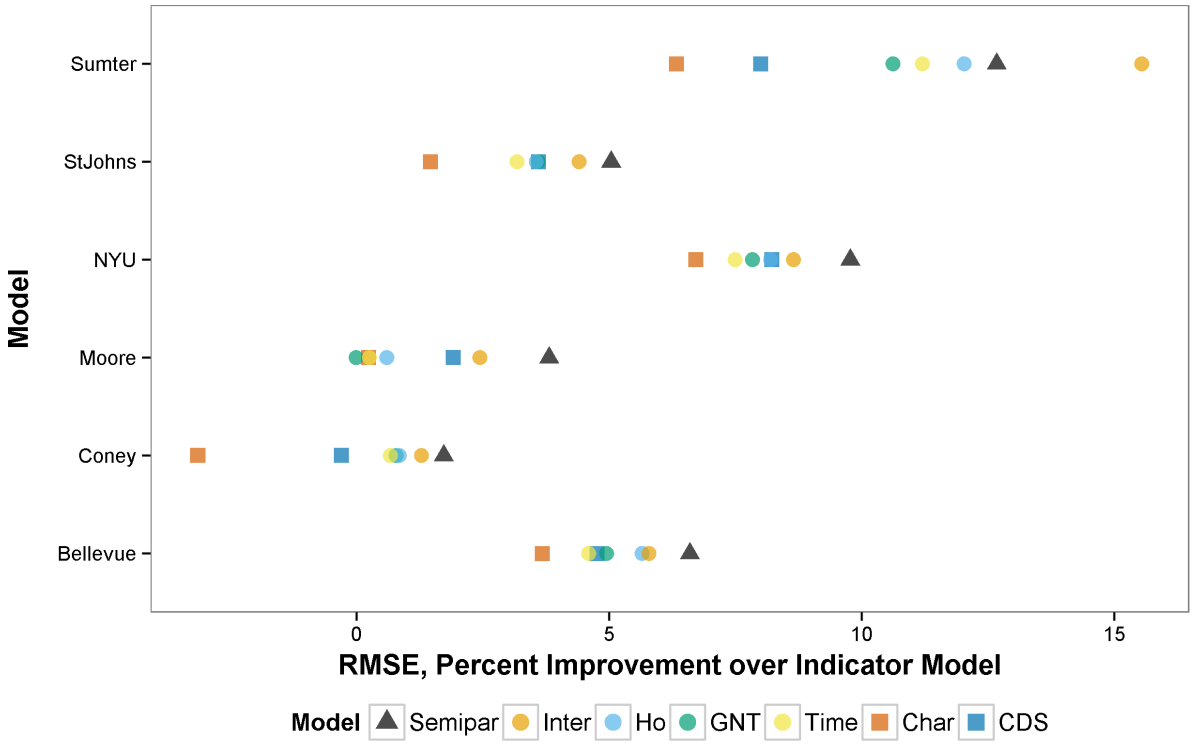


Figure 8 Relative Improvement in RMSE of Individual Predictions

Note: Improvement is Percentage Improvement in RMSE for each model over the *Indic* model.

The three other hospital effects models – *Ho*, *GNT*, and *Time* – tend to perform similarly, and are always clearly worse than *Semipar* and *Inter*. One of the pure characteristics models – *Char* – continues to perform worse than the hospital effects models, although it is better than *Indic*. *CDS*, the other characteristics model, is no longer clearly worse than the hospital effect models, and performs well for many of the service areas. For Coney and Moore, many of the models perform similarly or worse than our baseline *Indic*, which may indicate smaller differences in preferences across patients than in the other service areas.

In most situations, researchers will not have access to natural experiments like ours in order to assess models, but could use in-sample model performance to evaluate models. We examine whether in-sample performance can provide a good guide to out of sample performance in Figure 9. For each of the destroyed hospitals, we compare each model's performance for individual predictions in the period before the disaster to after the disaster. The blue line is the linear best-fit line across the models.

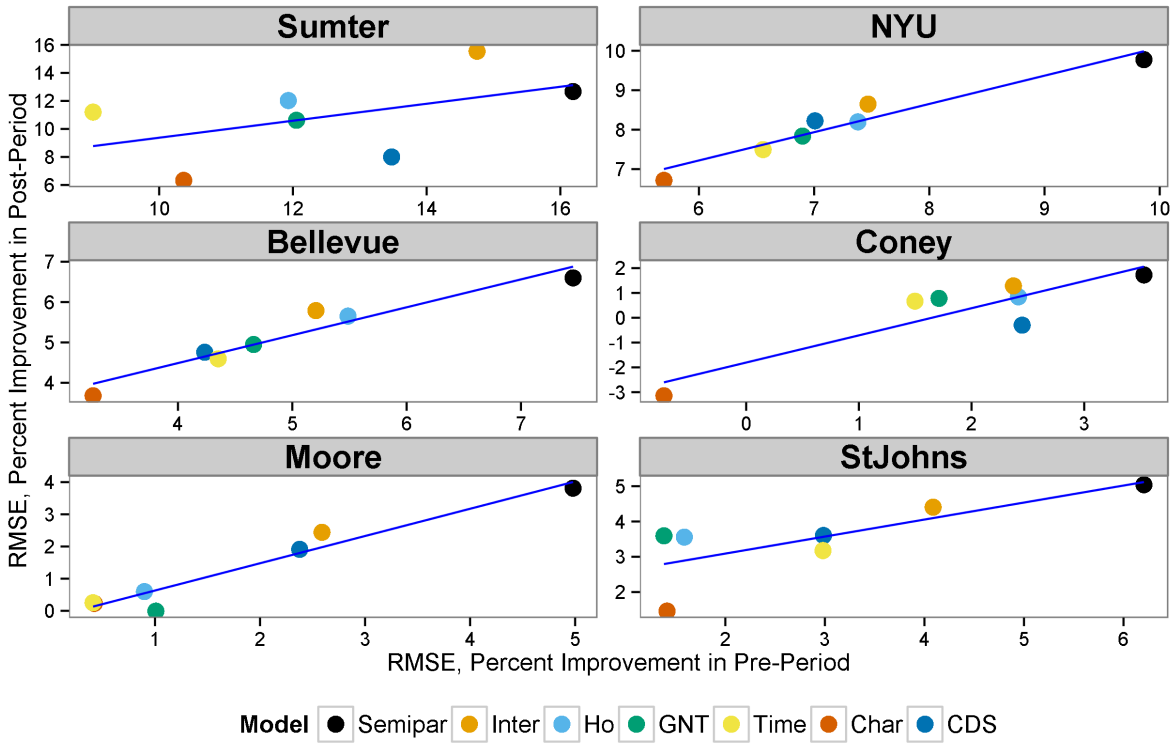


Figure 9 Relative Improvement in RMSE of Individual Predictions, Post-Period vs. Pre-Period

Note: Improvement is Percentage Improvement in RMSE for each model over the *Indic* model.

Overall, the performance of the model before the disaster is a good guide to its performance afterwards; all of the linear relationships are upward sloping and most models are close to the linear prediction. The models do tend to do worse compared to *Indic* in the post-period than they did in the pre-period. In addition, *CDS* appears to over-fit for Sumter and Coney, as it is better than *Ho*, *GNT*, and *Time* in the period before the disaster but

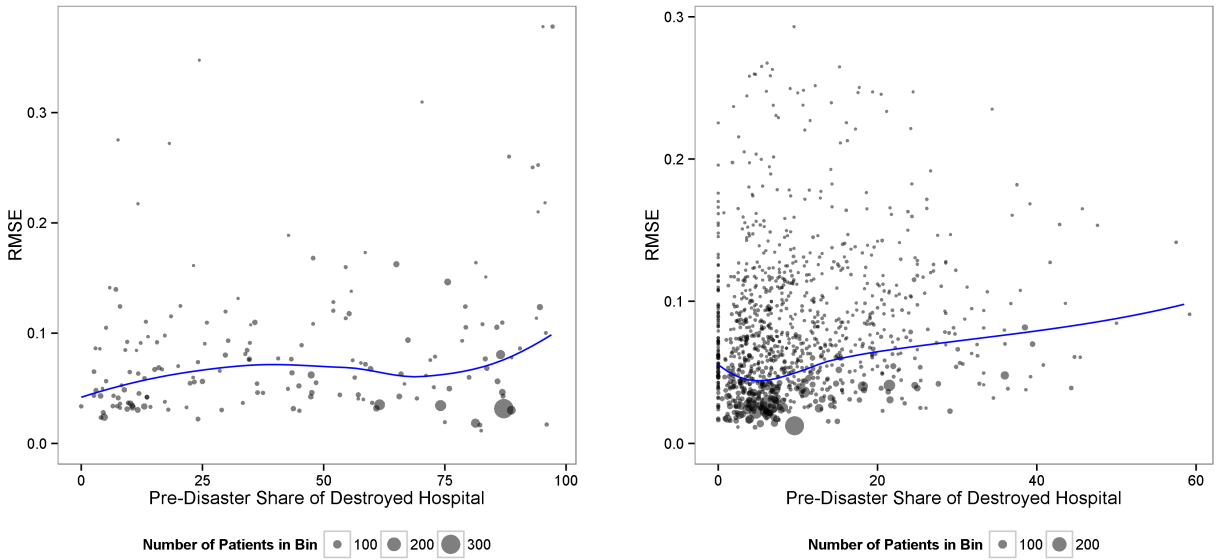
worse afterwards. Thus, in general model performance before the disaster predicts model performance after the disaster, but some models, such as *CDS*, may overfit the data.

4.2 Prediction Under a Changing Environment

The above results demonstrate that, on average, model flexibility generally trumps power considerations. However, it is possible that this is because we include in our estimates many patients for whom the choice set did not change after the disaster. If their preferred hospital was unaffected by the disaster, then the destruction of a non-preferred hospital should have no impact on their choices. The greater the number of patients included in our calculations that prefer a non-destroyed hospital, the more our out-of-sample validation resembles more traditional "split-the-sample" validation. In that environment, the flexibility may reflect a type of overfitting that delivers good predictions in the existing choice environment, but fails at extrapolations out of that environment.

We use two methods to focus on the patients who were more likely to experience the elimination of their preferred hospital following the natural disaster: a) patients whose characteristics place them in bins with a greater share of discharges from the destroyed hospital in the pre-disaster period and b) patients who used the destroyed hospital in the pre-disaster period.

For the first approach, we calculate the RMSE for each bin produced by *Semipar* and examine how bin level performance varies by the bin's share of the destroyed hospital for all of the models; Figure 10 depicts this relationship for Sumter and NYU. The size of each point is proportional to the number of patients in each bin; the blue solid line is the loess trend weighting each bin by its number of patients. For Sumter, the average RMSE increases as the share of the destroyed hospital increases, flattens out, and then increases again. For NYU, the average RMSE about doubles when going from the lowest pre-disaster share to the



(a) Sumter (b) NYU

Figure 10 Bin Level RMSE by Destroyed Hospital Share for *Semipar* model

Note: Blue solid line is the loess trend, weighting each bin by its number of patients. Each point is the RMSE for a particular bin, and its size is proportional to the number of patients in each bin.

highest pre-disaster share. This pattern is intuitive – when there is a change in the choice environment, the models generally do not predict as well.

While we should not expect the models to perform as well when there is a change in the choice environment, some models may preform relatively better than others. Therefore, we examine the relative performance by plotting the loess trend from each model across the bins. Figure 11 depicts these graphs for each model for all of the hospitals. For the New York and California hospitals, there is a cutoff share for the destroyed hospital, ranging from 25 to 55 percent, after which a parametric model performs better than *Semipar*. At high shares of the destroyed hospital, *Semipar* performs worse than most of the parametric models. We find one model to always perform the best for Sumter and Moore; this model is *Semipar* for Moore and *Inter* for Sumter.

For our second approach to identify the patients most likely to have lost their preferred hospital, we define the set of affected patients by looking at all patients that visited the de-

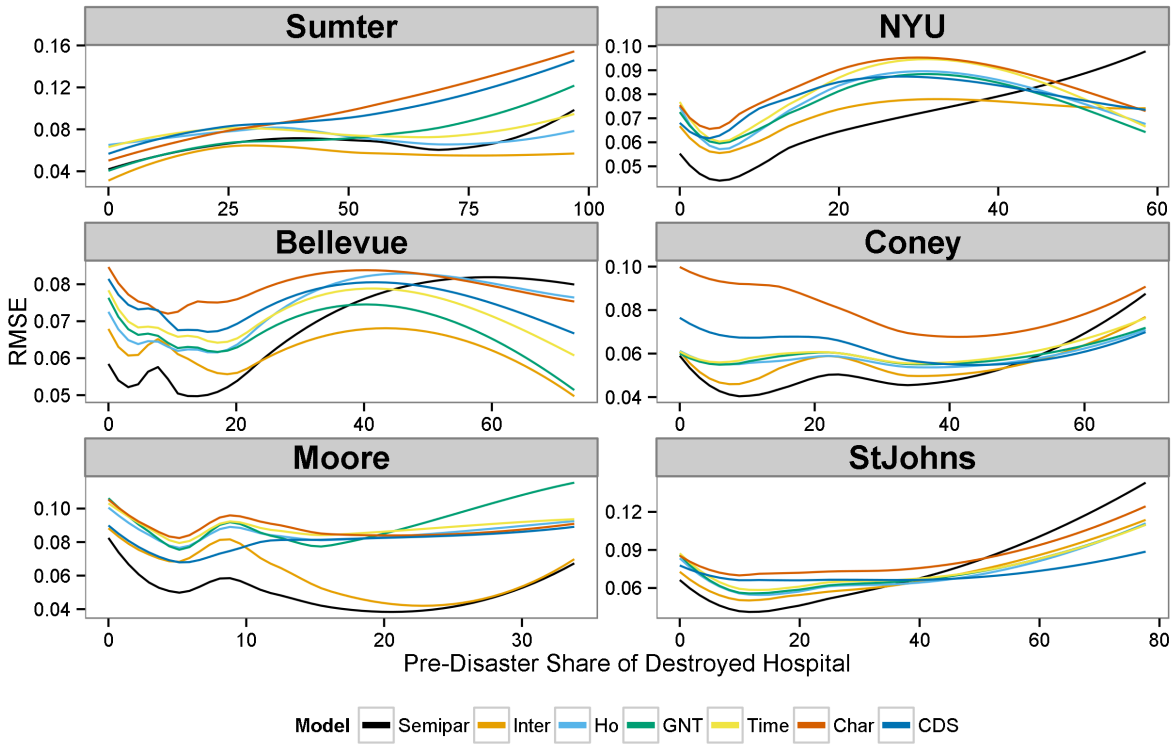


Figure 11 Bin Level RMSE by Destroyed Hospital Share for All Models

Note: Each line is the loess trend for a different model, weighting each bin by its number of patients in the pre-period.

stroyed hospital in the pre-disaster period. This approach allows us to focus on patients who are most likely to be affected in their facility choices by the destruction of the hospital, since they already received care at the destroyed hospital in the past.[12] We have patient identifiers for California and New York, and so use all patients who were admitted to the destroyed hospital in the year of the disaster or the previous two years prior for these disasters.[13]

Figure 12 displays the model performance at the individual level for these patients. For NYU, *Semipar* is the best model; for the other three hospitals, the best performing model

[12]Their continued preference for this hospital could either be the result of switching costs or time-invariant preferences for certain types of facilities; see Raval and Rosenbaum (2016) for a discussion of this issue. Which of these explanations is correct is irrelevant to our research question.

[13]For New York, we have to use a different dataset from HCUP, and so have to exclude patients discharged in 2013 as 2013 is not presently in our dataset. In accordance with our data use agreement for these data, we note that these data are from New York, State Inpatient Database (SID), Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality.

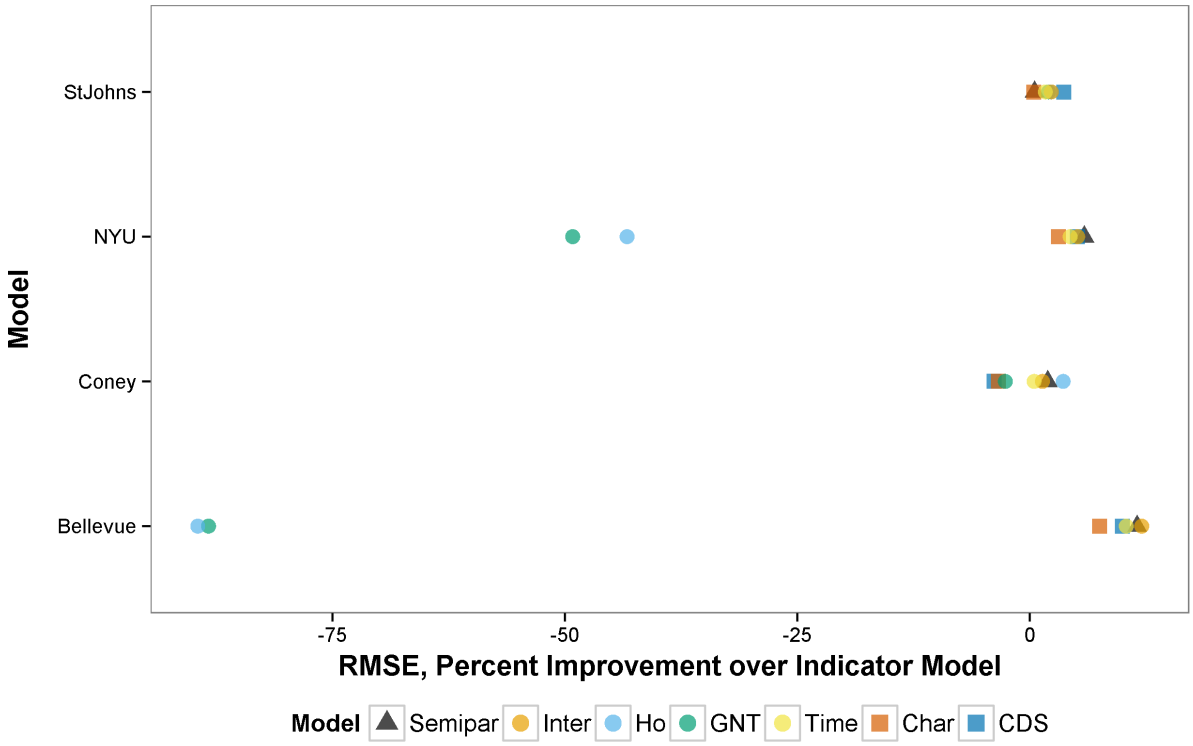


Figure 12 Relative Improvement in RMSE of Individual Predictions for Previous Patients of Destroyed Hospital

Note: Improvement is Percentage Improvement in RMSE for each model over the *Indic* model.

is also the best model at high shares of the destroyed hospital in Figure 11. *Semipar* still performs well for Coney and Bellevue, as it is the second best model in these cases, although it performs relatively poorly for St. John's. *Ho* and *GNT* perform very poorly for NYU and Bellevue.

Overall, the *Semipar* model has the best predictive ability when it is least likely that a patient's first choice hospital was destroyed. However, while *Semipar* still performs well, parametric models improve relative to the semiparametric approach when it is likely that a patient's first choice was eliminated. This decline in performance for *Semipar* could either be because of the reduced precision of the semiparametric model in those areas, or because a semiparametric approach is less effective in predicting following a change in the choice set.

For bins within the *Semipar* model in which the destroyed hospital has a large share,

choice probabilities will be based upon data from only a few individuals and so will have a high variance. To the extent that this is an issue, it would degrade the performance of *Semipar* relative to the other models. Therefore, it is particularly surprising that *Semipar* performs as well as it does in the population where the choice set changes.

4.3 Combining Multiple Models

So far, we have examined the performance of each model separately. However, since each model may rely on a different source of variation, a combination of models might provide better predictions than any given model separately. Further, the previous section shows suggestive evidence that different types of models provide more accurate predictions for different types of patients. These findings suggest that combining the predictions from multiple models may lead to better predictions of behavior than using a single "preferred model."

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 optimally combining macroeconomics forecasts (Timmermann (2006)). To apply the method to our context, we treat each patient as an observation, and regress the predictions from all the models on observed patient behavior. 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 many models will 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.

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 might receive some weight in the optimal model combination.

Formally, we 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:

$$y_{ih} = \beta^{Semipar}\hat{y}_{ih}^{Semipar} + ... + \beta^{CDS}\hat{y}_{ih}^{CDS} + \epsilon$$

where y_{ih} is the observed choice for patient i and hospital h and $\hat{y}_{ih}^{Semipar}$ is the predicted probability for patient i and hospital h for *Semipar*. We include the characteristics based models after re-estimating them including hospital indicators given the substantial bias in aggregate predictions demonstrated earlier.

Table IV Model Weights for Optimal Model Combination

Model	Sumter	Moore	NYU	Coney	Bellevue	StJohn's	Average
Semipar	0.21	0.61	0.52	0.59	0.56	0.54	0.50
CDS	0.13	0.23	0.48	0.10	0.28	0.34	0.26
Inter	0.57	0.16	0.00	0.16	0.00	0.07	0.16
Ho	0.09	0.00	0.00	0.15	0.16	0.00	0.07
Indic	0.00	0.00	0.00	0.00	0.00	0.04	0.01

Note: The second through seventh columns provide the model weights for the optimal model combination for each experiment's service area in the period after the disaster. Models not included in the table were given zero weight by the estimation. The last column provides the average weight for each model across the different experiments.

Table IV displays the model weights from these regressions for all models with positive weight for some experiment. We highlight two major findings. First, there is no one "preferred model." Within a given disaster, there is no single model that receives all of the weight; the largest weight any model receives is 61%. Across disasters, only *Semipar* and *CDS* (with hospital indicators) have positive model weights for all of the experiments. *Inter* receives a substantial amount of weight for Sumter, and receives some weight in three other experiments, while *Ho* receives positive weight in three experiments. While the importance of "robustness" checks in empirical work is well known, the positive contributions

of many model types in prediction illustrates that it is important to ensure that results are qualitatively similar for a variety of different specifications.

Second, the results from the model combination suggest that researchers should adopt an approach that puts roughly equal weight on a semiparametric and on a flexible parametric approach. On average, *Semipar* receives 50 percent of the weight, and receives a majority of the weight for all of the service areas except *Sumter*. The three parametric models receive the remainder at 26 percent for *CDS*, 16 percent for *Inter*, and 7 percent for *Ho*.

A model combination incorporating semiparametric and parametric models can be done informally, by estimating both models to ensure the consistency of results, and formally, by combining predictions from the models using a weighting of 50% on a semiparametric approach and 50% on a flexible parametric one. While there is variation across settings in the exact weighting and which parametric approach to use, our results strongly suggest that a flexible semiparametric specification is valuable in capturing variation that is missed by a parametric approach.

4.4 Absolute Performance

So far, we have examined the performance of the models relative to each other. But how well can the models predict actual choices, shares, and diversion ratios? We examine the absolute performance of discrete choice models by examining the RMSE between actual and predicted values. Figure 13a and Figure 14a display the RMSE for each hospital for aggregate shares, aggregate diversion ratios, and individual choice predictions for the different models for the *Semipar* and *Inter* models, respectively.

For *Semipar*, the RMSE on predictions of aggregate shares is quite small: between 0.7 percent and 2.2 percent across the models. The RMSE on aggregate diversion ratios is substantially higher, lying between 4 and 12 percent across the hospitals. The RMSE at the

individual level ranges between 19 and 27 percent across the hospitals.[14] The results from *Inter* are quite similar.

The RMSE can be large because the variance of actual and predicted choices is large, or because the correlation between predictions and observed choices is low.[15] To separate the effects of higher variance from lower correlation on the RMSE, we display the correlation coefficient between actual and predicted in Figure 13b and Figure 14b. At the aggregate level, the correlation coefficient is above 0.98 for all of the hospitals for *Semipar*, so the RMSE is likely due to the high variance of observed and predicted choices. The correlation coefficient for aggregate diversion ratios has a wide range across disasters, from 0.95 for Bellevue and 0.99 for Sumter to 0.65 for St. John's; only for Sumter and Bellevue are these close to the correlation of aggregate share predictions.

For *Semipar*, the correlation coefficient at the individual level is 0.64 for Sumter, between 0.42 and 0.48 for NYU, Bellevue, and St. John's, and about 0.35 for Coney and Moore. *Semipar* does substantially worse at individual prediction than aggregate prediction. However, given the complexity of modeling individual patient choice, individual predictions are still reasonably correlated with actual choices. Again, results are similar for *Inter*.

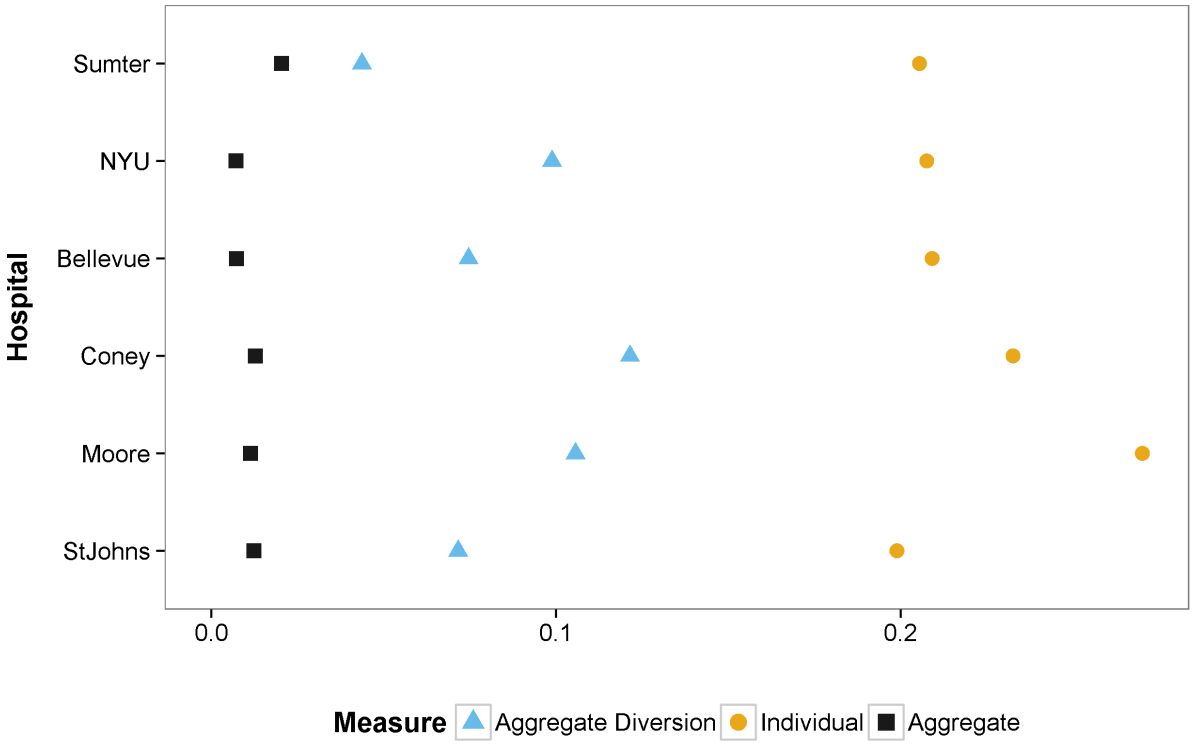
For both aggregate shares and aggregate diversions, our models performed the worst for St. John's. In Figure 15, we plot the time series of predictions for *Semipar*, *Inter*, and *Ho* against observed shares for two hospitals with large diversions – UCLA Medical Center

[14]The explained variance statistic, also known as Efron's pseudo R^2 (Efron (1978)), provides a transformation of RMSE that measures the amount of the overall variance in choices explained by the model. It is defined as:

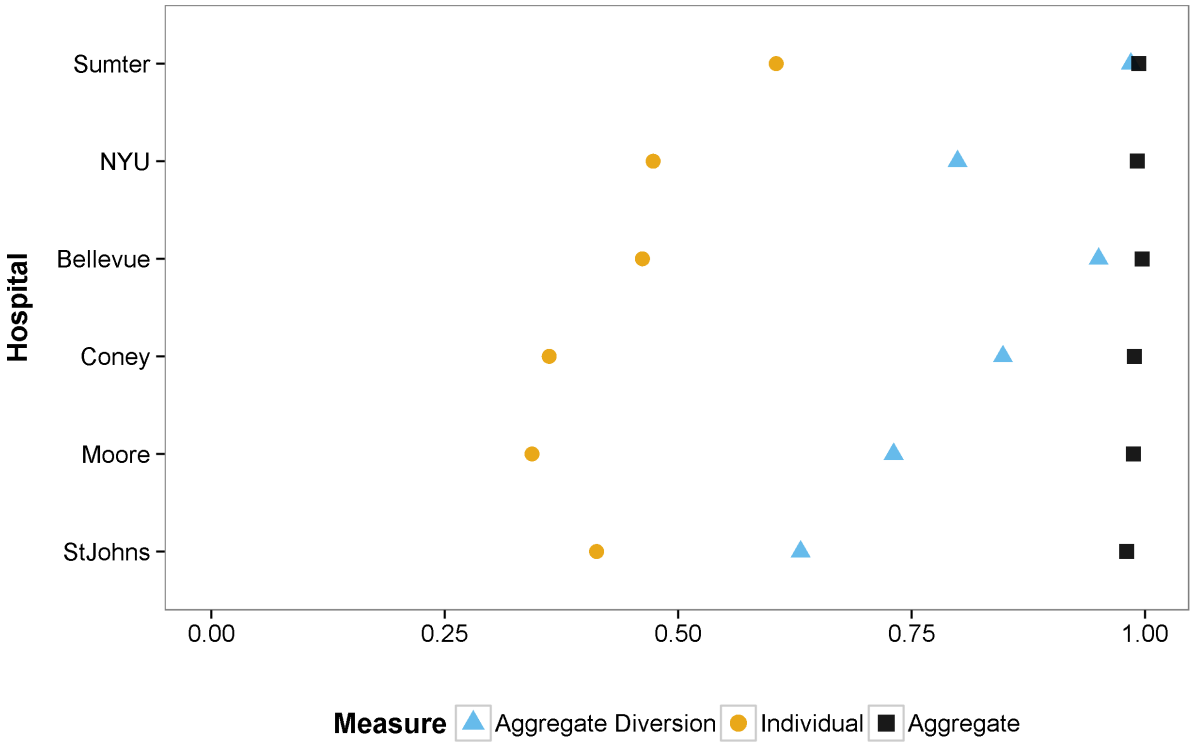
$$Explained\,Variance = 1 - \frac{\frac{1}{N}\sum_j [s_j - \hat{s}_j]^2}{\frac{1}{N}\sum_j [s_j - \bar{s}_j]^2}$$

The explained variance is between 96 to 99 percent for aggregate shares. For aggregate diversions, the explained variance ranges from 92 percent for Sumter, 80 percent for Bellevue, 70 percent for Coney, 57 percent for NYU, 44 percent for Moore, and 39 percent for St. John's. For individual choices, it is 40 percent for Sumter and between 12 to 23 percent for the other disasters.

[15]Formally, $E((y-\hat{y})^2) = V(y) + V(\hat{y}) - 2Cor(\hat{y},y)\sqrt{V(y)+V(\hat{y})}$ where y is the actual choice, $\hat{y}$ the prediction, Cor a correlation, V a variance, and $E((y-\hat{y})^2)$ the MSE across predictions. If the variance of observed values and predictions is equal, this simplifies to $2V(y)(1-Cor(\hat{y},y))$.



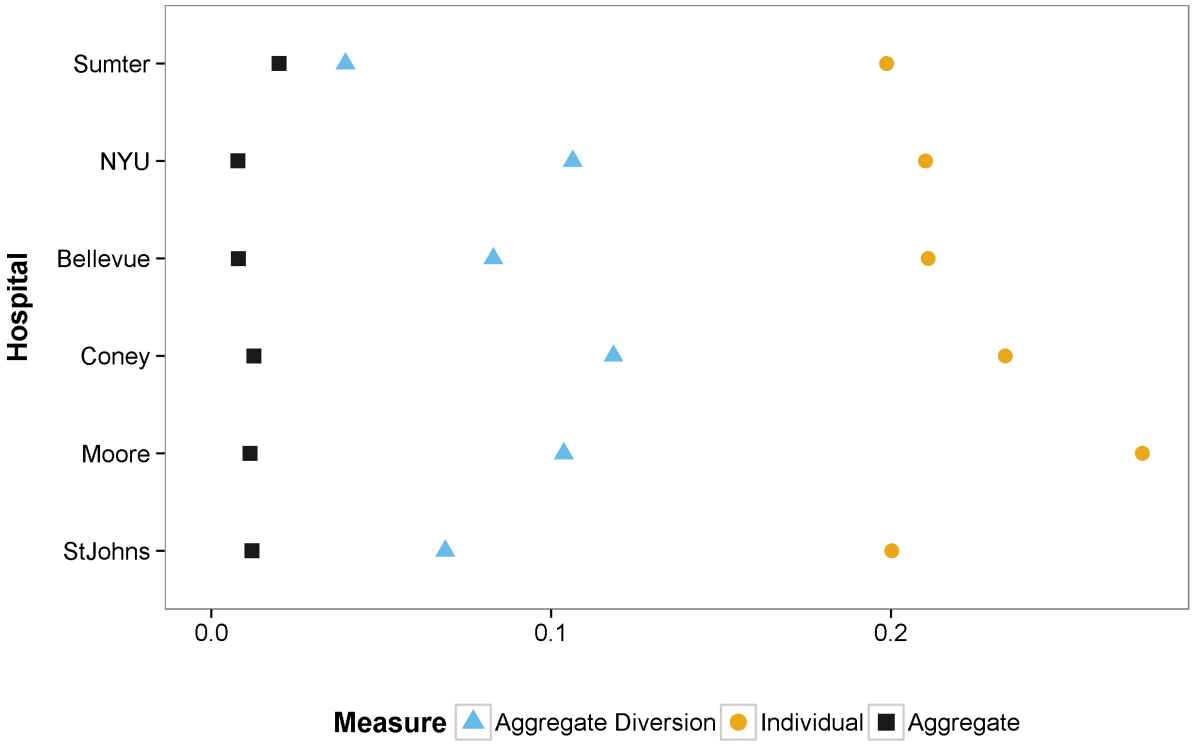
(a) RMSE



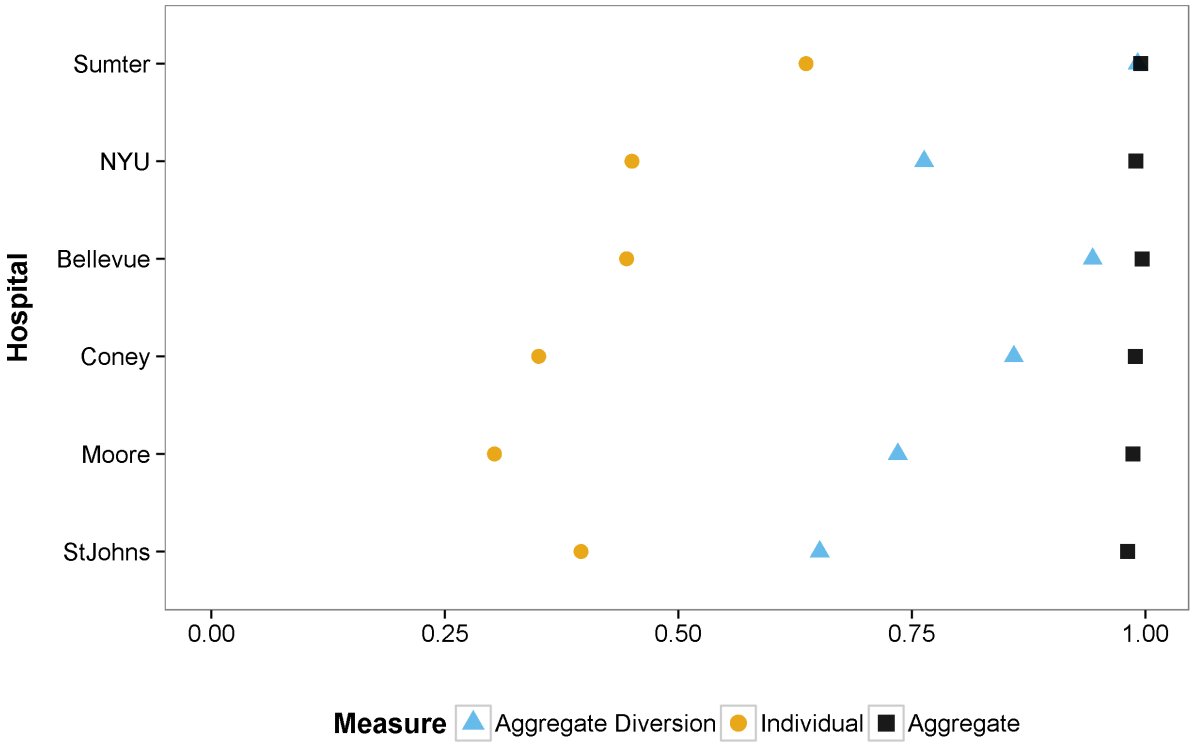
(b) Correlation Coefficient

Figure 13 RMSE and Correlation Coefficient for All Hospitals: Semipar Model

Note: The correlation coefficient is between the predicted values from the model combination model and observed choices or share in the period after each disaster.



(a) RMSE



(b) Correlation Coefficient

Figure 14 RMSE and Correlation Coefficient for All Hospitals: Inter Model

Note: The correlation coefficient is between the predicted values from the model combination model and observed choices or share in the period after each disaster.

and Santa Monica Hospital.[16] All of the models predict a big increase for Santa Monica Hospital, which occurs although it takes longer than the models predict. This may be because Santa Monica Hospital was also damaged in the earthquake, although it remained open with reduced capacity, and so it took longer for the hospital to fully reopen. UCLA's share rises from about 11 percent to 18 percent. None of the models predict the extent of the diversion to UCLA, although *Ho* performs better than the other two models. One explanation for this rise is that UCLA was making strategic investments in the Santa Monica area after the disaster. UCLA buys Santa Monica Hospital about a year after the disaster after merger talks between St. John's and Santa Monica Hospital break down.[17]

5 Policy Counterfactuals

So far, we have examined the predictive accuracy of a number of different models, and shown that flexible models such as *Semipar* and *Inter* perform better than the others for individual prediction. It is not clear, however, whether the differences in model fit are large enough to impact policy counterfactuals of interest. We examine this through an application to mergers, using the framework of Capps et al. (2003) to consider the connection between model accuracy and expected changes in consumer welfare.[18]

Conditional on accurately modeling δ, the logit distributional assumption makes it straightforward to assess the welfare consequences of alterations to patients' choice sets. A patient's

[16] The predictions for *Ho* change dramatically between 1992 and 1993 because many of the *Ho* interactions are based on AHA variables that change at the yearly level.

[17] See http://articles.latimes.com/1994-12-17/business/fi-9940_1_santa-monica-hospital-medical-center for a discussion of these changes.

[18] Interestingly, all four areas hit by disasters had substantial merger activity in the year following the disaster; we saw evidence of this for California in the previous section with the merger of Santa Monica Hospital and UCLA Medical Center. In Georgia, Sumter Regional, the destroyed hospital, merged with Phoebe Putney, which, as we saw in Figure 6, was the hospital with the biggest diversion after the disaster. In New York, Beth Israel Medical Center broke off merger talks with NYU Langone and merged with Mount Sinai. New York Downtown Hospital, located close to NYU and Bellevue, merged with New York Presbyterian. Finally, in Oklahoma, Norman Regional Health System, the owner of Moore Medical Center, merged with a major physician group in the area.

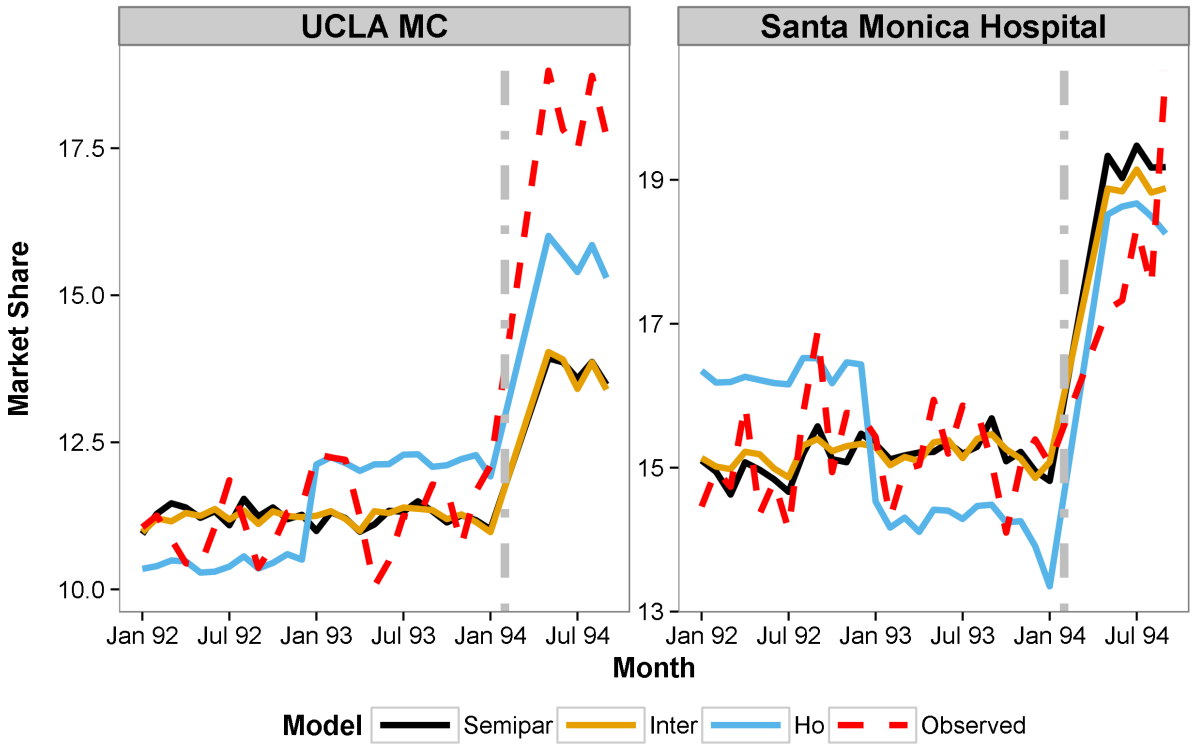


Figure 15 Aggregate Market Shares, Predicted and Observed, for St John's

Note: Red dashed line is the observed series of market shares. The grey vertical dot-dash line depicts the quarter of the disaster. February through April 1994 are omitted from the plot, since that is the period immediately following the earthquake that we omit from our post-period.

ex-ante consumer surplus from being able to access a hospital, which we henceforth follow Capps et al. (2003) and others in calling the patient's "willingness to pay" (WTP) for the hopsital, is defined for a given hospital as the change in utility for that patient if the hospital is excluded from the choice set. A patient's WTP for a hospital k given choice set H is the difference between the patient's ex-ante utility for hospital choice set H and their utility for the choice set after hospital k is dropped from the choice set:

$$WTP_i(k)(H) = W_i(H) - W_i(H/k) = \log\left(\frac{1}{1 - s_{ik}(H)}\right)$$

where s_{ik} is the share of hospital k for patient of type i given choice set H.[19] To construct the

[19]For convenience, we drop the dependence on H in our notation.

overall WTP for a hospital, one simply integrates $WTP_i(k)$ over the distribution of patients.

For hospital merger simulations and screens, economists are often interested in how WTP changes at the provider system level after the merger. If hospitals are close substitutes, the WTP of a system composed of the two hospitals together is greater than the sum of the WTP of each hospital separately. The WTP of the combination of two hospitals k and l is:

$$WTP_i((k,l)) = W_i(S) - W_i(S/(k,l)) = \log\left(\frac{1}{1 - s_{ik} - s_{il}}\right)$$

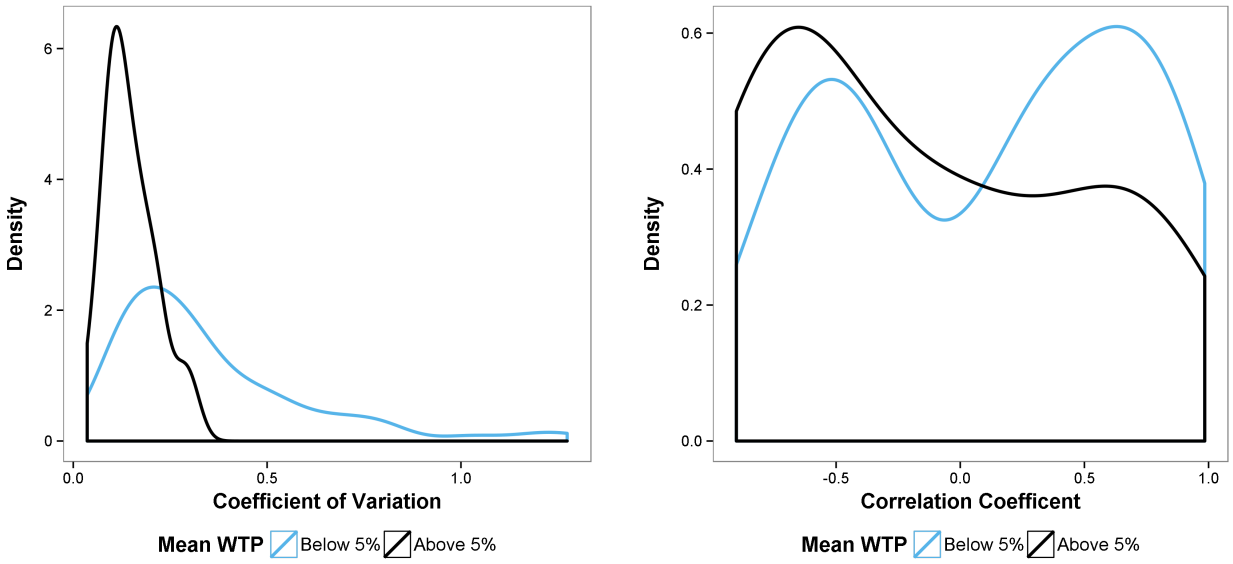
The post-merger change in WTP for an individual is the difference between the WTP for the combined hospital system and the individual hospitals, and so can be estimated as:

$$\begin{aligned} \Delta WTP_i((k,l)) &= WTP_i(S/(k,l)) - WTP_i(S/k) - WTP_i(S/l) && (3) \\ &= \log\left(\frac{1}{1 - s_{ik} - s_{il}}\right) - \log\left(\frac{1}{1 - s_{ik}}\right) - \log\left(\frac{1}{1 - s_{il}}\right) && (4) \end{aligned}$$

At the aggregate level, the change in willingness to pay is the WTP for the combination of k and l minus the sum of the WTP for k and the WTP for l. The overall change in WTP is thus the integral over this quantity; we report the percentage change in WTP, which is the integral over $WTP_i(k,l)$ divided by the integral over the sum of $WTP_i(k)$ and $WTP_i(l)$.

We employ the WTP framework to examine the effect of counterfactual mergers in each destroyed hospital's market. In order to explore a range of different types of mergers, we simulate a merger of the destroyed hospital with every other hospital in its service area. This produces a total of 95 counterfactual mergers. We assess the connection of model performance to policy outcomes by dividing our simulated mergers into two groups based on their percent change in WTP, which can heuristically be associated with expected price changes following a merger.[20] Specifically, we define two sets of simulated mergers based on

[20]See, e.g., Robert Town's address to the American Bar Association, http://apps.americanbar.org/antitrust/at-committees/at-hcic/pdf/past-programs/20100601_town.pdf. In that address, Town



(a) Coefficient of Variation

(b) Correlation Coefficient

Figure 16 Density Estimates from Counterfactual Merger Predictions

Note: Each figure depicts smoothed kernel density estimates from estimates of the coefficient of variation for the percent change of WTP and the correlation coefficient between individual level RMSE and the percent change of WTP across all simulated hospital mergers.

whether or not they have an average percent change in WTP of above 5%, with those having a higher change in WTP being more likely to result in substantial competitive harm.

The left figure of Figure 16 shows a density plot of the coefficient of variation of the percent change in WTP for the two groups of hypothetical mergers. The coefficient of variation, the standard deviation divided by the mean percent change in WTP, is computed using different econometric models within a single hypothetical merger.[21] This mimics how

suggests using a 10% change in WTP as a merger screen, stating that "In modestly concentrated markets, a 10% increase in the WTP for any hospital is noteworthy.". This is broadly consistent with the published and public evidence correlating changes in WTP to changes in price. For example, Capps et al. (2003) found that WTP increases in the range of 12% to 20% were associated with predicted price increases of 10% to 12%. In the ProMedica case, Robert Town as the FTCs expert estimated a model that predicted that the 13.5% increase in WTP would lead to 16.2% higher prices. For a useful summary of this evidence, see Capps (2014, p. 472). Moreover, the figures reported in published work are broadly consistent with those reported in working papers. In particular, identifying the correlation between changes in WTP and changes in price cross-sectionally using a set of consummated mergers, Garmon (2016) estimates a slightly higher average elasticity of around 1.

[21] For example, for a given hypothetical merger suppose we computed percent change in WTP of 3, 6, and 9. The coefficient of variation would be 0.5 (3, the standard deviation, divided by 6, the mean).

different analysts may use different econometric models to analyze the same transaction in a merger investigation. We then pool the coefficient of variation from the hypothetical mergers across all of the natural disasters to make the density plot.

The figure illustrates the significant relative variation in the predicted harm for mergers with a relatively small percent change in WTP and the much smaller relative variation in harm for mergers with larger predicted harm. This is reassuring. In the cases where the percentage change in WTP results suggest a substantial reduction in competition, there is much less variation than in cases where it is unlikely to be relevant for merger enforcement.

Nevertheless, the coefficient of variation is 0.15 on average for mergers with an average change in WTP of over 5 percent, a potentially substantial magnitude. For example, suppose that a given merger has a predicted 10% change in WTP on average across models and a coefficient of variation of 0.15. Predicted changes ranging from a 7% change in WTP to a 13% change in WTP woud lie within 2 standard deviations of the mean.

While the analysis of the connection between model consistency and economic significance tells us that using different models could lead to different policy conclusions, it is important to know the types of mistakes that will be made from using different types of models. To study this, we compute the correlation between out-of-sample RMSE and percent change in WTP for all of the models within each hypothetical merger. In the right figure of Figure 16, we show a density plot of these correlations. As we did above, we split the density plot between mergers that are more and less likely to result in competitive harm. A negative correlation suggests that more accurate models predict higher harm, while a positive correlation suggests the opposite.

For mergers with low estimated changes in WTP, we find that the distributions of these correlations across mergers is bimodal. For these mergers, it is hard to detect a pattern in the bias from using a less accurate model – sometimes the merger effect will be overstated and sometimes understated. For 22% of those mergers there is a negative correlation between

percentage change in WTP and RMSE with a magnitude greater than 0.5, while for 33% there is a positive correlation with magnitude greater than 0.5.

However, when looking at mergers with a percent change in WTP of above 5%, where scrutiny is more likely, more accurate models generally predict higher harm. For 60% of such mergers, there is a negative correlation between percent change in WTP and RMSE. Moreover, for 42% of mergers with a percent change in WTP of above 5%, the negative correlation is greater than 0.5 in magnitude. While not conclusive, this general trend suggests that the use of less accurate models could make it more difficult for enforcers to make correct enforcement decisions.

When would better fitting models typically produce higher changes in WTP? One answer is that the type of unobserved heterogeneity that the best fitting models are capturing has more patients with high probabilities for both merging hospitals. In Appendix E, we show formally that the percent change in WTP will be larger on average for probability distributions that increase the share of patients with high probabilities for both hospitals and low probabilities for both hospitals, keeping the marginal distributions of probabilities the same.[22] Thus, if more flexible models like *Semipar* and *Inter* better represent high probabilities for both merging hospitals for certain patients given the same aggregate shares, they will also predict larger changes in WTP from a merger.

6 Robustness

In computing our main results, we assume that individuals are able to choose any hospital within their region and that the distribution of hospital patients is unaffected by the natural disaster. The first assumption would be violated if patients faced restricted health insurance

[22]The change in WTP, equation (4), is a supermodular function. This supermodularity means that patients with high probabilities of visiting both merging hospitals will have large changes in WTP for the merged hospital system. The aggregate change in WTP will then rise when the share of patients with high probabilities for both hospitals rises, even if aggregate shares for each hospital remains the same.

networks or hospital capacity constraints. The second would be violated if many people moved in the aftermath of the natural disaster. In this section we argue that our results are robust to addressing concerns surrounding both of these assumptions. We outline our robustness results here and include more detail in Appendix D.

Commercially insured patients have typically had largely unconstrained choices of facilities. For example, Ho (2009) reports that 83 percent of hospitals and plans reach agreement. However, to ensure that our results are robust to the possibility that some commercially insured individuals face constrained choice sets due to being in "narrow network" plans, we reestimate all of the models only using Medicare patients. While this population is a narrower demographic group than our full sample, and a significantly smaller sample, they should have unrestricted access to all hospitals in the area. Using this population, we find qualitatively similar results to our baseline results, although *Semipar* tends to perform worse at the individual level.

To address concerns that the hospitals faced capacity constraints following the disaster, we approximate a measure of hospital capacity using our data. Using these measures, we find that five hospitals are potentially impacted by capacity constraints following the disaster, four of which are in areas of Brooklyn close to Coney Island. For three out of these five hospitals we underpredict share and for two we predict correctly. If capacity constraints were severely impacting our results, we would expect to systematically overpredict the shares of the constrained hospitals.

Another concern following a disaster is that the population shifted in such a way that the pre-disaster admissions are not a good proxy for the post-disaster admissions. For Coney, St. John's, and Sumter, we examine our models' performance excluding the areas that were most affected by the disaster. Using this population, we find qualitatively similar findings to our baseline specifications.

We also examine how the case mix changed after the disaster in a number of dimensions,

including age, diagnosis, diagnosis acuity, payer type, and number of admissions. The main substantial change we find are falls in that the total number of admissions per month falls by 6 to 14 percent across service areas, which is consistent with the evidence of falling admissions after hospital exits in Petek (2016). This may reflect an extensive margin for hospital admissions. In addition, the fraction of patients under 18 falls substantially for Moore and Sumter. We also examine our model predictions separately for cardiac and labor/pregnancy patients, two groups of people for which the extensive margin may be less relevant. We find qualitatively similar results for these patients as for the overall sample.

One reason that models may do a poor job of prediction is that patients' choices following the disasters are driven by where the destroyed hospital's physicians practice. We examine this theory for the New York hospitals, and find that it is unlikely to be a concern in our case for two reasons. Following the disasters, the physicians from the destroyed hospitals saw many fewer patients than the average from the previous months; for Bellevue and Coney the decline was above 90% and for NYU it was about 60%. Second, the regular patients of the destroyed hospitals typically went to different hospitals after the disaster than the doctors did; the physician and patient diversion ratios are uncorrelated with each other.

7 Conclusion

For many economists, the experiment is the benchmark by which to judge empirical economic research (Angrist and Pischke, 2008). However, experiments and quasi-experiments typically can only focus on a narrow population and set of market conditions. For example, it is not possible to predict the counterfactual impact of a merger of two firms using an experiment that merges them, and there may be no quasi-experiment of similar firms to the merging ones. Further, it is not possible to assess welfare in the absence of a model of consumer utility, and utility parameters must be estimated under assumptions on preferences. Therefore, many

economists use structural modeling to predict policy counterfactuals over a much broader set of events than focusing on quasi-experiments would allow.

In this paper, we take a similar approach as LaLonde (1986) did for program evaluation, and compare the results obtained from structural models to quasi-experiments. Using detailed micro data on patient choice of hospital facility, we compare consumer substitution patterns obtained from structural discrete choice models of demand to those obtained from an exogenous change in consumers' choice sets. We construct a laboratory in which to assess the performance of structural demand models, such that econometricians can use these models in other markets with greater confidence in their conclusions.

Our qualitative conclusions are robust across markets. First, we find that the best performing models allow for substantial preference heterogeneity. Second, flexible specifications for consumer heterogeneity frequently involve a bias-variance tradeoff. Therefore, we suggest that researchers consider the robustness of their conclusions to parametric and semi-parametric approaches that balance this tradeoff differently. Finally, differences in model specification can lead to qualitatively different predictions of competitive effects. While this is not surprising, it emphasizes the importance of model selection and testing.

Where possible, we hope to see more studies that are able to use experimental analyses to help cross-validate structural modeling. In industries where this is possible, this type of external validation could become standard practice. As Nevo and Whinston (2010) remind us, "in general, structural analysis and credible identification are complements."

References

Ackerberg, Daniel, C. Lanier Benkard, Steven Berry, and Ariel Pakes, "Econometric Tools for Analyzing Market Outcomes," *Handbook of Econometrics*, 2007, *6*, 4171–4276.

Aguirregabiria, Victor, "Empirical Industrial Organization: Models, Methods, and Applications," 2011.

Angrist, Joshua D. and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, 2008.

_ **and Jorn-Steffen Pischke**, "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics," *The Journal of Economic Perspectives*, 2010, *24* (2), 3–30.

Bayer, Patrick, Ferreira Fernando, and Robert McMillan, "A Unified Framework for Measuring Preferences for Schools and Neighborhoods," *Journal of Political Economy*, 2007, *115.*

Berry, Steven, James Levinsohn, and Ariel Pakes, "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market," *Journal of Political Economy*, 2004, *112* (1), 68–105.

Bresnahan, Timothy, "The Apple-Cinnamon Cheerios War: Valuing New Goods, Identifying Market Power, and Economic Measurement," 1997.

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

Capps, Cory S, "From Rockford to Joplin and Back Again: The Impact of Economics on Hospital Merger Enforcement," *The Antitrust Bulletin*, 2014, *59* (3), 443–478.

Carlson, Julie A., Leemore S. Dafny, Beth A. Freeborn, Pauline M. Ippolito, and Brett W. Wendling, "Economics at the FTC: Physician Acquisitions, Standard Essential Patents, and Accuracy of Credit Reporting," *Review of Industrial Organization*, 2013, *43* (4), 303–326.

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.

Conlon, Christopher T. and Julie Holland Mortimer, "An Experimental Approach to Merger Evaluation," *NBER Working Paper*, 2013.

Dafny, Leemore, Kate Ho, and Robin S. Lee, "The Price Effects of Cross-Market Hospital Mergers," *NBER Working Paper Series*, 2016, *22106.*

Efron, Brad, "Regression and ANOVA with zero-one data: Measures of residual variation," *Journal of the American Statistical Association*, 1978, *73*, 113–121.

Einav, Liran and Jonathan Levin, "Economics in the Age of Big Data," *Science*, 2014, *346.*

Garmon, Christopher, "The Accuracy of Hospital Merger Screening Methods," *mimeo*, 2016.

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.

Goldberg, P.K., "Product differentiation and oligopoly in international markets: The case of the US automobile industry," *Econometrica: Journal of the Econometric Society*, 1995, pp. 891–951.

Goolsbee, Austan and Amil Petrin, "The Consumer Gains from Direct Broadcast Satellites and the Competition with Cable TV," *Econometrica*, 2004, *72* (2), 351–381.

Gordon, Brett and Wesley Hartmann, "Advertising Effects in Presidential Elections," *Marketing Science*, 2013, pp. 19–35.

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.

Hausman, Jerry A, "Valuation of New Goods under Perfect and Imperfect Competition," in "Bresnahan, Timothy and Gordon, Robert J.," University of Chicago Press, 1997, pp. 207–248.

Hendel, Igal and Aviv Nevo, "Measuring the Implications of Sales and Consumer Inventory Behavior," *Econometrica*, November 2006, *74* (6), 1637–1673.

Ho, Kate and Robin Lee, "Insurer Competition in Health Care Markets," *mimeo*, 2015.

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

_ , "Insurer-Provider Networks in the Medical Care Market," *The American Economic Review*, 2009, *99* (1), 393–430.

LaLonde, Robert J., "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *The American Economic Review*, 1986, pp. 604–620.

Lancaster, Kelvin J., "A New Approach to Consumer Theory," *The Journal of Political Economy*, 1966, pp. 132–157.

May, Sean M., "How Well Does Willingness-to-Pay Predict the Price Effects of Hospital Mergers?," *mimeo*, 2013.

McFadden, Daniel, "Econometric Models of Probabilistic Choice," in Daniel McFadden and Charles F. Manski, eds., *Structural Analysis of Discrete Data and Econometric Applications*, Cambridge: The MIT Press, 1981.

_ **, 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.

Nevo, Aviv and Michael D. Whinston, "Taking the Dogma out of Econometrics: Structural Modeling and Credible Inference," *The Journal of Economic Perspectives*, 2010, pp. 69–81.

Pathak, Parag A. and Peng Shi, "Demand Modeling, Forecasting, and Counterfactuals, Part I," Technical Report, National Bureau of Economic Research 2014.

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

Raval, Devesh and Ted Rosenbaum, "What Really Drives Spatial Demand? Evidence from Hospital Choice," *mimeo*, 2016.

_ , _ , and Steven A. Tenn, "A Semiparametric Discrete Choice Model: An Application to Hospital Mergers," *mimeo*, 2015.

Shepard, Mark, "Hospital Network Competition and Adverse Selection: Evidence from the Massachusetts Health Insurance Exchange," 2016.

Tchen, Andre H., "Inequalities for Distributions with Given Marginals," *The Annals of Probability*, 1980, *8* (4), 814–827.

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

Todd, Petra E. and Kenneth I. Wolpin, "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility," *The American Economic Review*, 2006, pp. 1384–1417.

A Disaster Timelines

In this section, we give a 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.[23] Over 1,600 housing units in Santa Monica alone were damaged with a total cost of $70 million.[24]

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.[25] By the same time, many of those with damaged houses had found new housing.[26]

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.[27] 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).[28]

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 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.[29] 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

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

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

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

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

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

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

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

later.[30]

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.[31]

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).[32] Major bridges reopen on October 30th and NYC schools reopen on November 5th.[33] 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.[34] 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 and surgical intensive care. The maternity unit and pediatrics reopened on January 14th, 2013. [35] 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.[36]

Bellevue Hospital Center reopened limited outpatient services on November 19th, 2012.[37] However, Bellevue dis not fully reopen inpatient services until February 7th, 2013.[38] Coney Island Hospital opened an urgent care center by December 3, 2012, but patients were not admitted inpatient.

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

[31] 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.

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

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

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

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

[36] 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.

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

[38] See

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.[39]

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.[40] Maps of the tornado's path demonstrate that while some areas were severely damaged, nearby areas were relatively unaffected.[41]

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 with a tentative opening of fall 2016.[42]

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 them on the period after the disaster. We omit the month of the disaster from either period, excluding anyone either admitted or discharged in the disaster month. The length of the pre-period and post-period in general depends upon the length of the discharge data that we have available.

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

[39] 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/.

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

[41] 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.

[42] See https://www.normanregional.com/en/locations.html?location_list=2 and http://kfor.com/2013/11/20/moore-medical-center-destroyed-in-tornado-to-reopen-in-december/.

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	NA	NA

C Model Details

In this section, we give a narrative description of each of the models we test. In Appendix F, we show a detailed table of included variables across all models.

Capps, Dranove, and Satterthwaite *(CDS)*

In one of the earliest applications of discrete choice models to the hospital choice literature, Capps et al. (2003) suggest that the different hospitals can be modeled exclusively in characteristic space (Aguirregabiria, 2011). In other words, they assume a utility function that includes a rich set of interaction terms between patient attributes (age, sex, income, etc.) and hospital characteristics (for-profit status, teaching hospital, nursing intensity, etc.), as well as a measure of the travel time from the patient's home to the hospital. They also include several interactions between hospital services and patient disease characteristics; for example, between a patient's admission for childbirth and the presence of a labor and delivery room at a hospital. As a result, their specification does

not include any time invariant controls for individual facilities. We attempt to implement as close a version of the model used in Capps et al. (2003) as we can using our data.

Characteristic *(Char)*

Other models than just *CDS* have relied exclusively on interactions between hospital and patient characteristics. We include one additional such model *Char* that was used by Garmon (2016). It includes a different, and smaller, set of such interactions than *CDS*.[43]

Hospital Indicators *(Indic)*

This model is composed of only time-invariant hospital indicator variables.

Hospital Indicators and Travel Time *(Time)*

This model includes hospital characteristics, time, and squared time; May (2013) suggests that this model performs just as well as more sophisticated models.

Ho *(Ho)*

This model attempts to mirror the specification used in Ho (2006). The specification integrates more elements of the characteristics approach of *CDS* into *Time* by including a fairly rich set of interactions between patient and hospital characteristics. For example, the model explicitly incorporates the possibility that oncology (or other types of) patients may be disproportionately more interested in receiving care at teaching hospitals, while also allowing some teaching hospitals to be consistently more desirable than others. It also includes a set of hospital indicators.

Gowrisankaran, Nevo, and Town *(GNT)*

In a recent paper, Gowrisankaran et al. (2015) include a different set of interactions than Ho (2006). In particular, they add in a set of interactions between the acuity "weight" assigned to the DRG of the patient and the hospital indicators. Such controls help to account for the possibility that the payoffs to visiting some hospitals are highly dependent on the rareness or severity of the complication. They also include several interactions between patient characteristics and travel time, and a couple of interactions between patient characteristics and hospital characteristics.[44] We attempt to as closely as possible replicate their specification.

[43] Our model is based on an earlier version of this paper. In the more recent version of the paper, there is an additional indicator for whether a patient was admitted through the emergency room that we have not included.

[44] In addition, Gowrisankaran et al. (2015) use a specification that includes the copay owed by the patient. Consistent with the general "option demand" hypothesis (Gaynor et al., 2013) that consumers are largely indifferent to prices when selecting among those providers in their insurance network, the estimated coefficient on this term was of very small economic magnitude (albeit statistically significant).

Weights and Other Interactions *(Inter)*

This model builds on both *Ho* and *GNT*, focusing on the possibility of important unobserved heterogeneity affecting the relative desirability of different hospitals to different consumers. This model, addresses the possibility of such unobservables by including an even richer set of interaction terms. In particular, it builds on *GNT* by incorporating interactions between key major diagnosis categories (e.g., oncology, cardiology, etc.) and each hospital indicator, travel time and each hospital indicator, as well as between DRG weight and hospital indicator. This new set of variables explicitly accounts for the possibility that some hospitals are particularly specialized in certain clinical categories. It also includes interactions of travel time with patient characteristics.

Semiparametric Bin Model *(Semipar)*

This model, detailed in Raval et al. (2015), takes even more seriously the possibility of unobserved heterogeneity affecting the relative desirability of different hospitals for different patients. It does this by permitting individuals of different ages, medical categories, condition severities, and zip codes to have arbitrarily different preferences for hospitals. For example, zip code might reflect further patient demographics than captured in the discharge data that affect hospital preferences, and therefore this heterogeneity would not be captured by simply including a "distance" term in the the estimation.

This approach pools together individuals with common characteristics and estimates the choice probabilities for that group. All individuals in that group are assumed to have the same ex-ante choice probabilities. As discussed in Carlson et al. (2013) and Raval et al. (2015), this highly flexible approach is actually extremely computationally efficient despite being equivalent to the inclusion of a very large number of indicator variables in a multinomial logit model.

In this particular implementation, we use an iterative procedure to generate predicted probabilities. Our initial approach is to use bins based on zip code, acuity group, age group, major MDC, in that order. If the resulting bin is too thin – meaning that it contains less than 20 individuals, we drop a category. This minimum bin size functions analogously to an bandwidth parameter; we choose a minimum bin size of 20 because Raval et al. (2015) find that estimates of diversion ratios and willingness to pay are relatively insensitive to intermediate ranges of the minimum bin size.

For Online Publication

D Robustness

D.1 Medicare-only Sample

As a robustness check, we restrict the sample to only the Medicare population, estimate our models on this population, and then examine the performance of the models for this population after the disaster. For the states for which Fee for Service Medicare and Managed Care Medicare are distinguished, we exclude Managed Care Medicare as well. The Medicare sample should have unrestricted access across all of the hospitals in the choice set. The Medicare sample is also considerably smaller than the full dataset; this may hurt the performance of more flexible models such as *Semipar* because of power considerations. Finally, the Medicare sample has a different variety of treatment conditions than the full sample; for example, Medicare patients probably do not value the pediatric or obstetrics departments of a hospital.

Figure 17 and Figure 18 contain the aggregate and individual level performance results for the Medicare sample. The prediction results for the Medicare sample are, in general, quite similar to the overall sample. While *Semipar* does perform worse at the individual level, which is consistent with it losing power with a much smaller dataset, it remains one of the three best models for all of the disasters.

D.2 Sample Removing Destroyed Areas

We also conduct a robustness check 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.

For Sumter, we remove the two zip codes comprising the city of Americus; the destruction of the Americus tornado was concentrated in the city of Americus. For Coney Island, we remove three zip codes which had the most amount of damage after the disaster, as based on post-disaster claims to FEMA; these zip codes are on the Long Island Sound and so 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 procedure removes 9 zip codes, including all 5 zip codes in Santa Monica.[45]

[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/

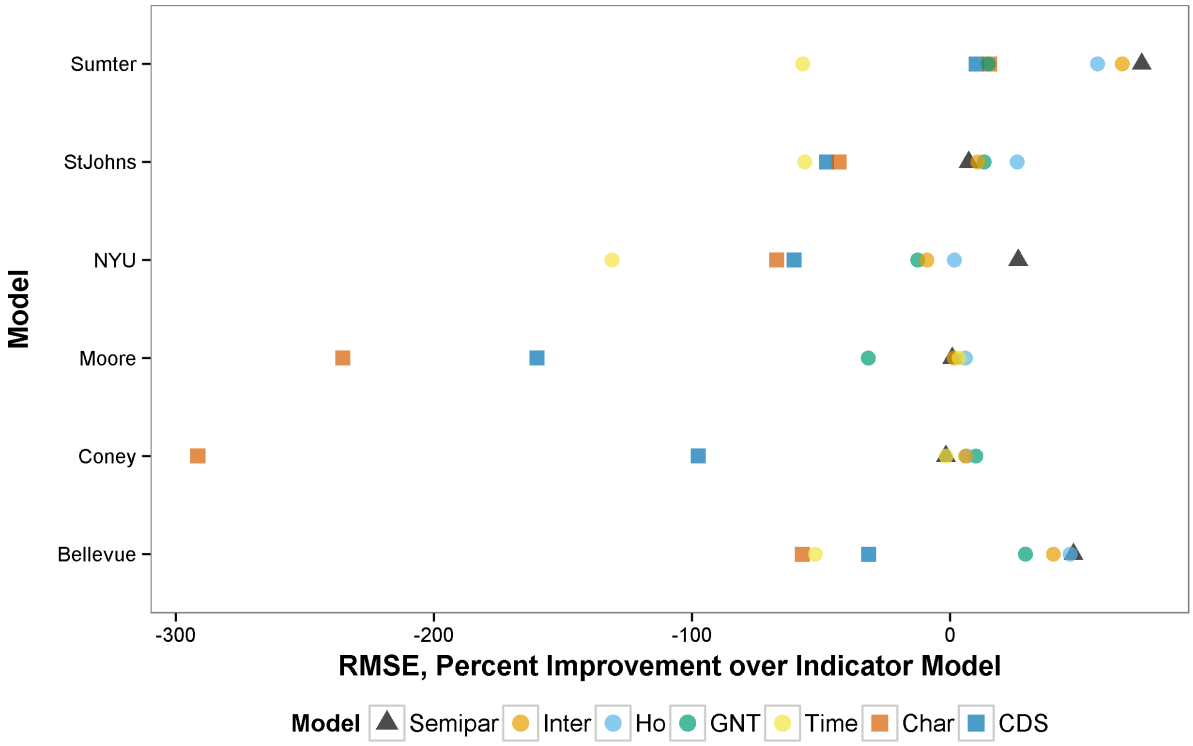


Figure 17 Relative Improvement in Predictive Accuracy at the Aggregate Level Across Hospitals: Medicare Sample

Note: Improvement is Percentage Improvement in RMSE for each model over the *Indic* model. Only patients admitted with Medicare insurance are included.

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.

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 then estimate the models on the full sample but restrict our performance validation measures to the restricted sample removing destroyed areas. Figure 19 and Figure 20 contain the aggregate and individual level performance results. We find very similar results to the full sample; the characteristics models tend to do poorly at aggregate prediction. For individual and aggregate prediction, *Semipar* and *Inter* are the best models, although at the aggregate level, all of the models tend to perform worse than *Indic* for Sumter.

`OES%20Reports/NR%20EQ%20Report_Part%20A.pdf`, Appendix C, for the Northridge MMI data by zip code.

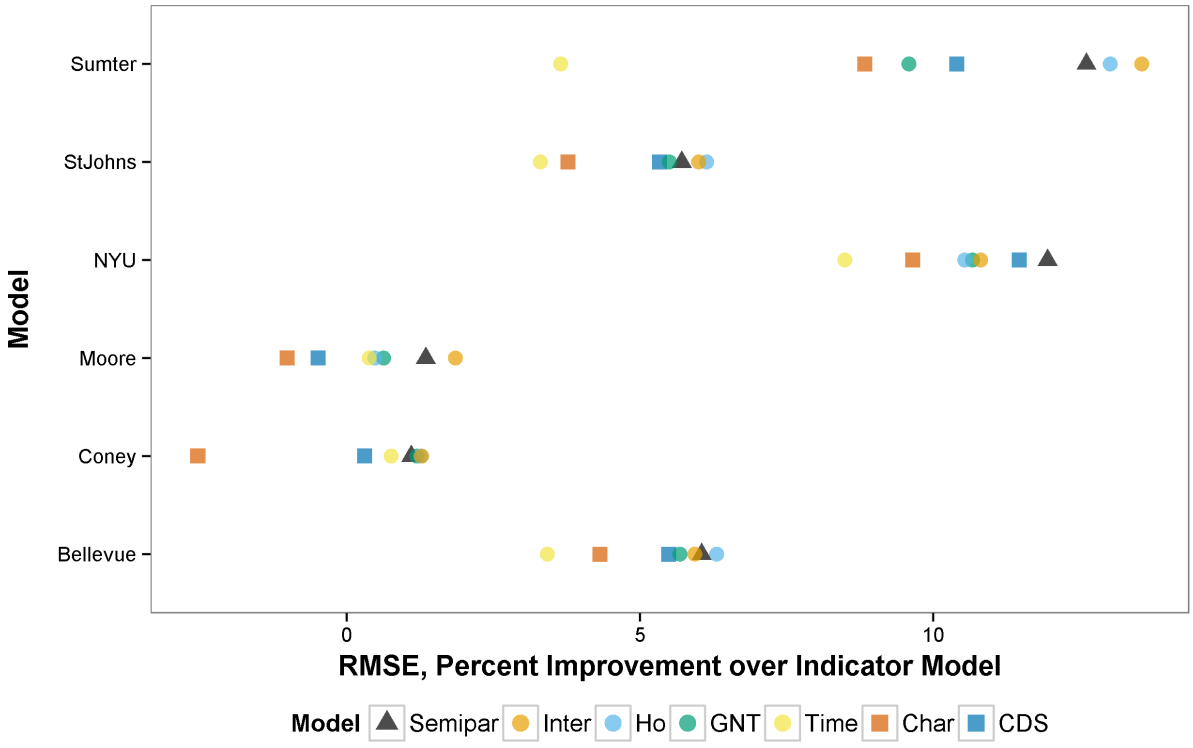


Figure 18 Relative Improvement in Predictive Accuracy at the Individual Level Across Hospitals: Medicare Sample

Note: Improvement is Percentage Improvement in RMSE for each model over the *Indic* model. Only patients admitted with Medicare insurance are included.

D.3 Capacity Constraints

One concern with the natural disaster experiments is that the remaining hospitals are capacity constrained after the disaster. Since the estimated models do not take into account capacity constraints, they might then overpredict diversion to capacity constrained hospitals, and so perform poorly.

For the Sumter and St. John's disasters, we have data on the date that each patient was admitted and discharged, and so can explicitly measure capacity for each day. For the Moore and Sandy disasters, we only have data on the month of admission and discharge. We thus calculate monthly capacity as a sum of each patient's length of stay for patients admitted in that month divided by the total number of days in the month. While crude, we compare this capacity measure to true capacity for the hospitals in the choice set for Sumter and find that it is approximately unbiased.

Defining "capacity constrained" as at least 90 percent capacity utilization after the disaster, none of the hospitals in the Sumter or Moore disaster are constrained. For the New York hospitals, several hospitals breach the capacity constraint of 90 percent at some point in the sample. However,

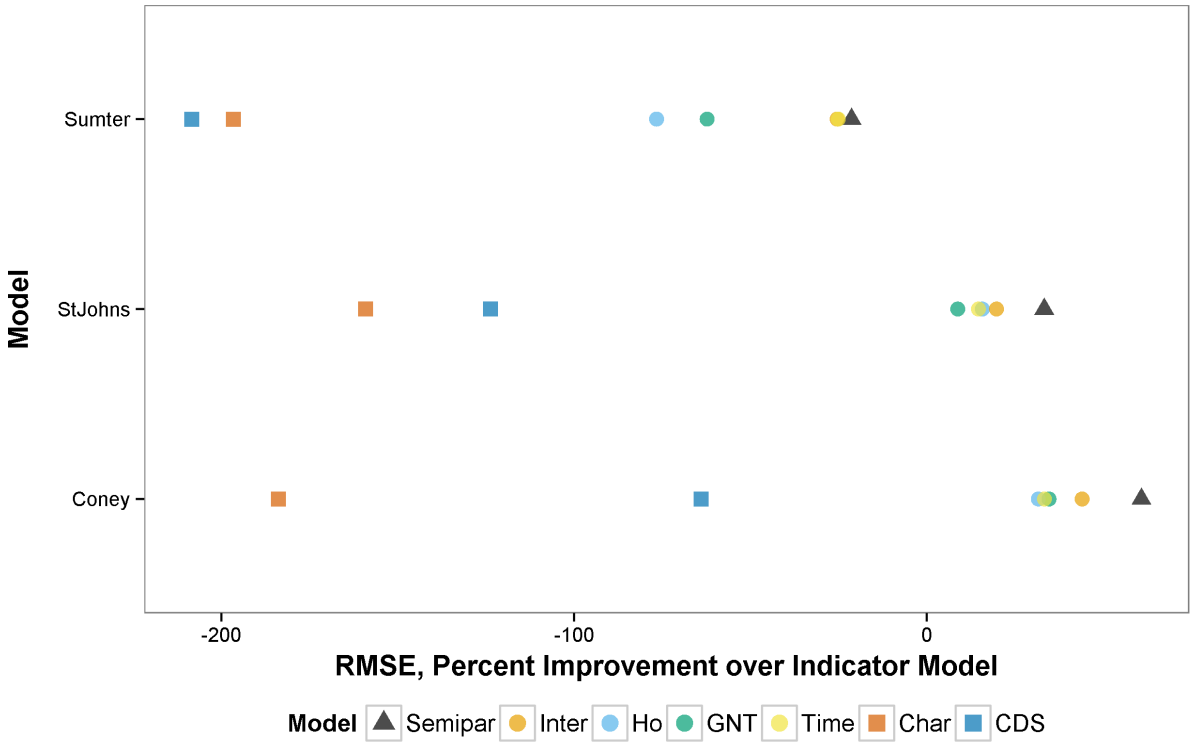


Figure 19 Relative Improvement in Predictive Accuracy at the Aggregate Level Across Hospitals: Removal Sample

Note: Improvement is Percentage Improvement in RMSE for each model over the *Indic* model. Zip codes with some destruction from the disaster removed.

five hospitals move from never constrained before the disaster to constrained in both months after the disaster: Kings County Hospital Center, Lutheran Medical Center, Maimonides Medical Center, New York Community Hospital, and New York Downtown Hospital. The first four hospitals are located in Brooklyn and are in the choice set of Coney Island; two are in NYU's choice set and one in Bellevue's choice set. New York Downtown Hospital is located in lower Manhattan and is in all of the choice sets.

Figure 21 depicts the aggregate market shares and our predictions based on our *Inter* and *Semipar* models for the four Brooklyn hospitals. We underpredict market shares for three of these hospitals and correctly predict one. If capacity constraints were seriously affecting model performance, we would expect to overpredict actual market shares for most of the hospitals. For New York Downtown Hospital, we predict the share of the hospital after the disaster roughly correctly in both the NYU and Bellevue experiments.

For California, Daniel Freeman Memorial Hospital is overcapacity both before and after the disaster. No hospital becomes capacity constrained after the disaster, including Santa Monica Hospital, whose capacity drops considerably due to disaster related damage.

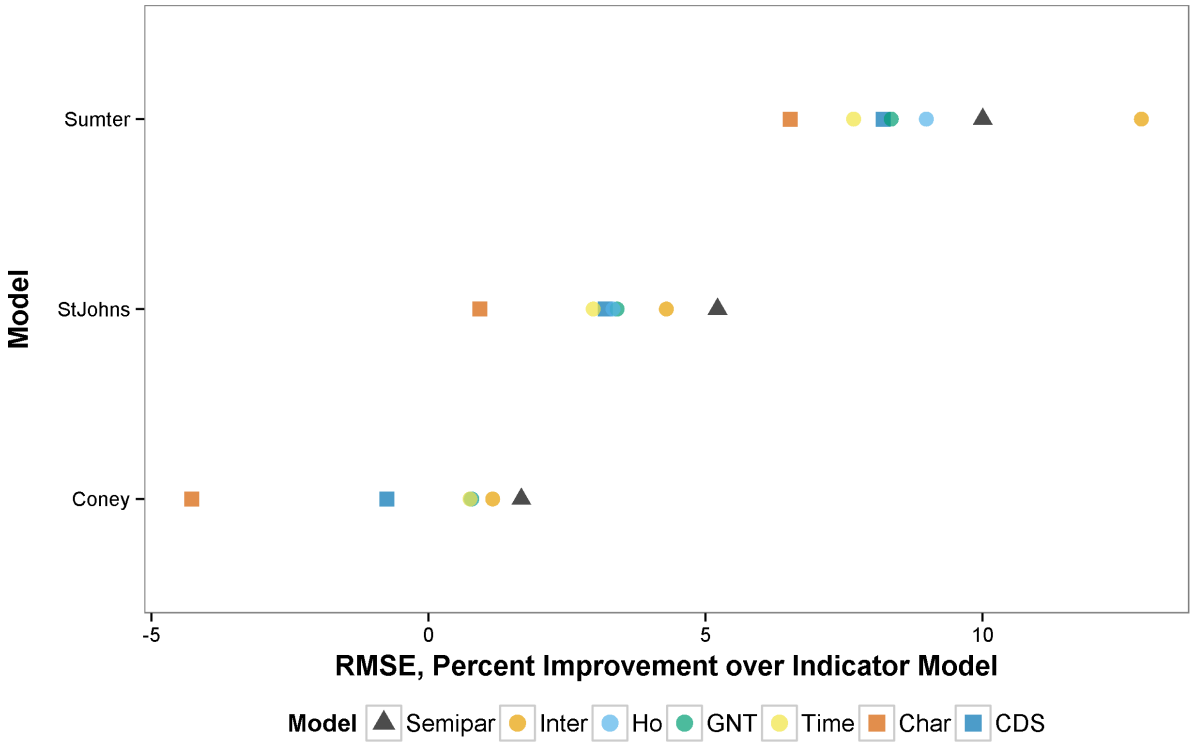


Figure 20 Relative Improvement in Predictive Accuracy at the Individual Level Across Hospitals: Removal Sample

Note: Improvement is Percentage Improvement in RMSE for each model over the *Indic* model. Zip codes with some destruction from the disaster removed.

D.4 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 would impair the performance of simpler models such as *Time* and *Indic* that do not control for patient characteristics; they may also indicate substantial changes to the service area that make the disaster less of an clean experiment.

In Table D-1 to Table D-6, 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.

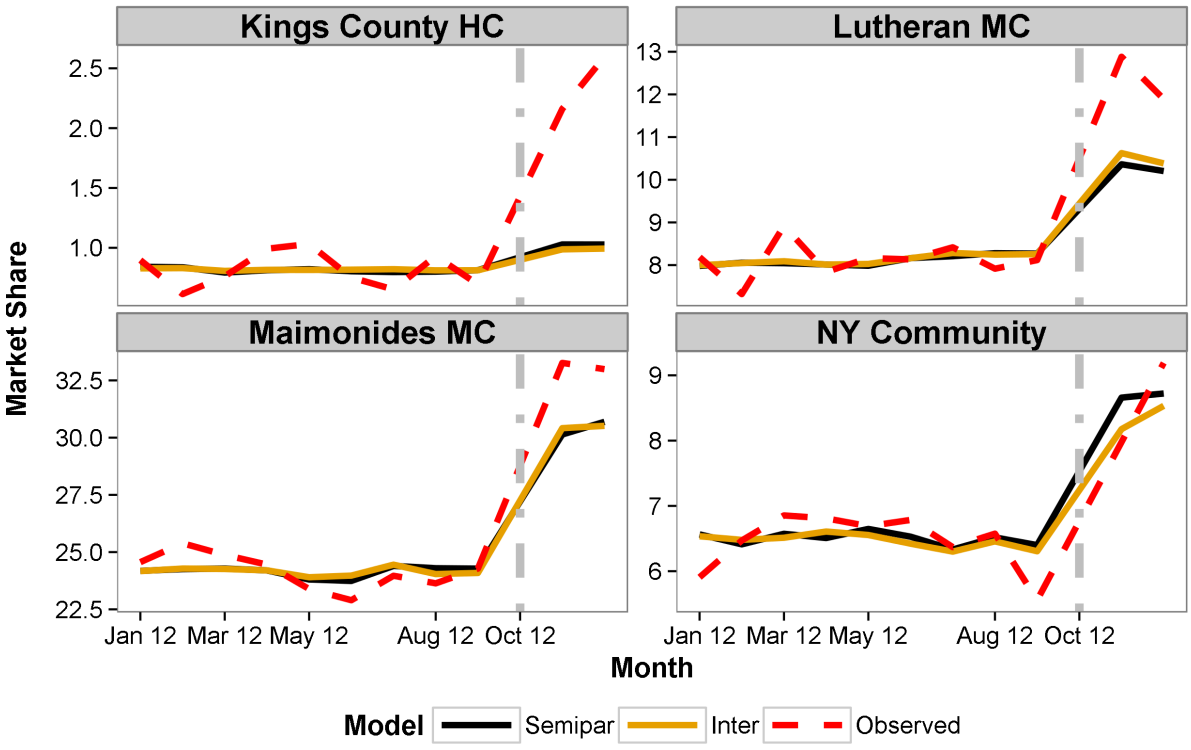


Figure 21 Aggregate Market Shares, Predicted and Observed, for Coney

Note: Red dashed line is the observed series of market shares. The grey vertical dot-dash line depicts the quarter of the disaster.

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 D-1 Changes in Case-Mix for Moore

Variable	Pre	Post	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-period, while the third column is the average of the variable in the post-period. The fourth column is the percent difference from the pre-period to the post-period.

Table D-2 Changes in Case-Mix for Coney

Variable	Pre	Post	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-period, while the third column is the average of the variable in the post-period. The fourth column is the percent difference from the pre-period to the post-period.

D.5 Physicians

In this section, we explore, and present evidence against, the possibility that patients' hospital choice following the disasters was primarily driven by where the hospitals' physicians practiced. Using the discharge data from New York, we look at where the doctors that generally practice in the destroyed hospitals admit patients in the months following the storm. We compare the behavior of the physicians to the post-storm behavior of the patients that frequently used the destroyed hospitals in the months and years preceding the disaster.

For this analysis, we look at the patients and doctors that were regular users of the destroyed hospitals in the months preceding their closure. We select physicians who are recorded as an operating physician for at least 50 patients in the first 9 months of 2012, and patients who had five or more admissions to one of these hospitals in the 2 years prior to the disaster and in the first 9

Table D-3 Changes in Case-Mix for NYU

Variable	Pre	Post	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-period, while the third column is the average of the variable in the post-period. The fourth column is the percent difference from the pre-period to the post-period.

Table D-4 Changes in Case-Mix for Bellevue

Variable	Pre	Post	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-period, while the third column is the average of the variable in the post-period. The fourth column is the percent difference from the pre-period to the post-period.

months of 2012.

We compute a "diversion-ratio" illustrating the change in usage patterns for patients and physicians from the period prior to the storm to the period after. For the patients that were regular users of a given hospital, we compute the difference in the share of admissions to *other* area hospitals in the first 9 months of 2012, prior to Superstorm Sandy, to the last two months of 2012, after the hospital was destroyed. We scale that change in the share of admissions by the share of admissions from these patients to the destroyed hospital in the pre-disaster period.

We do a similar calculation for the physicians. For physicians that were regular users of the destroyed hospital, we compute their share of admissions. These shares are computed by using all admissions where these physicians are listed as the operating physician. Similar to the calculation with patients, we compute the difference in the share of admissions to *other* area hospitals in the first 9 months of 2012, prior to Superstorm Sandy, to the last two months of 2012, after the hospital

Table D-5 Changes in Case-Mix for St. Johns

Variable	Pre	Post	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-period, while the third column is the average of the variable in the post-period. The fourth column is the percent difference from the pre-period to the post-period.

Table D-6 Changes in Case-Mix for Sumter

Variable	Pre	Post	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-period, while the third column is the average of the variable in the post-period. The fourth column is the percent difference from the pre-period to the post-period.

was destroyed. We scale that change in share of admission by the share of admissions from these physicians to the destroyed hospital in the pre-disaster period.

These calculations provide multiple reasons to doubt that patients were following their doctors in their post-disaster choice of hospital. Many of the physicians that were regular doctors at the destroyed hospital saw many fewer patients than their typical load in the months following the disaster. Their level of admissions fell by 58% for doctors at NYU, 92% for doctors at Bellevue, and almost 100% for doctors at Coney. Among those physicians that did admit patients, Figure 22 illustrates that there is essentially no correlation between patient and physician diversion ratios. This suggests that hospitals where the regular doctors of the destroyed hospitals went were not necessarily the hospitals where the regular patients went. Further, there is suggestive evidence that patterns here were consistent with the the ownership of the hospital. For example, Bellevue Hospital is a flagship hospital of the the public New York Health and Hospitals Corporation ("HHC"). Those

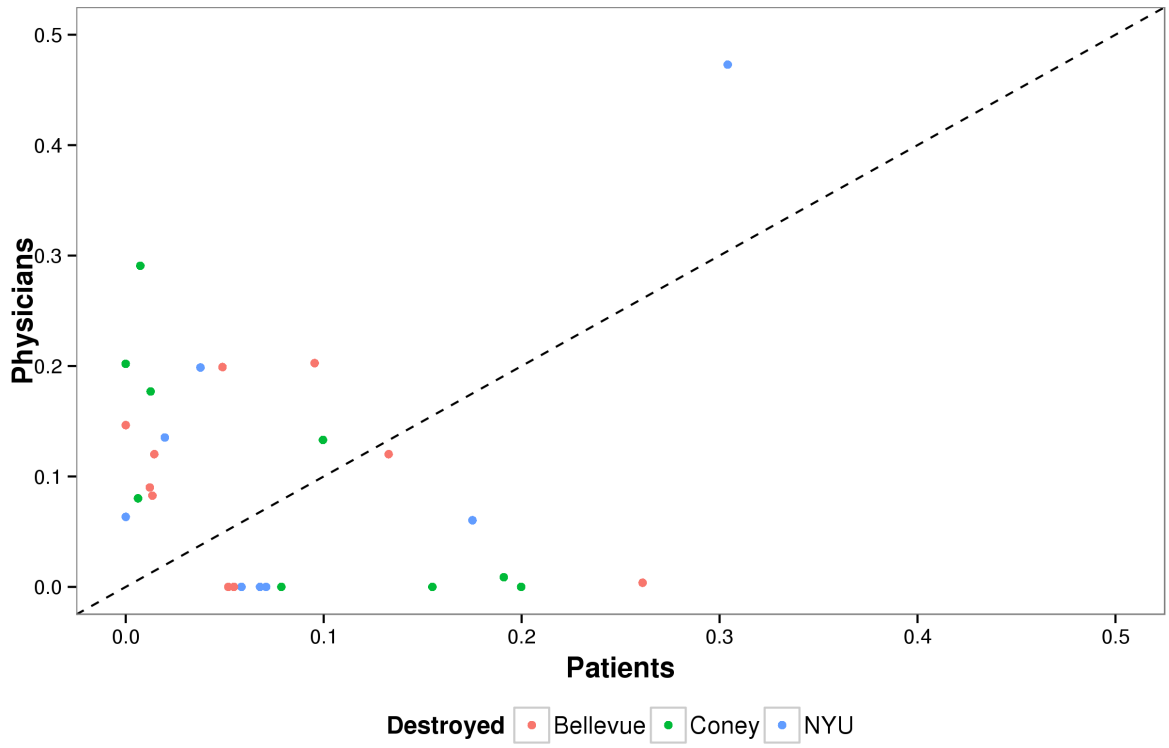


Figure 22 Diversion Ratios, Patients vs. Physicians

Note: This graph shows any hospital with above a 5% diversion of either patients or physicians. Each dot represents a pair of "diversion ratios" (for patients and physicians) from the destroyed hospital to a given non-destroyed hospital.

Bellevue physicians that continued operating on patients most often went to other hospitals in the HHC. In contrast, regular Bellevue patients most often went to Beth Israel Hospital, which is only approximately 10 blocks south.

D.6 Different Diagnosis Categories

So far, we have examined the performance of each choice model over patients across all patients. In this section, we examine two important classes of patients based on their diagnosis: cardiac patients (with a Major Diagnostic Category of 5) and obstetrics patients (with a Major Diagnostic Category of 14). We estimate the models on all patients, but then separately examine their performance for patients in each of the two diagnosis categories. Figure 23 and Figure 24 contain the results for cardiac and obstetrics patients at the individual level. In both cases, the results are fairly similar; *Semipar* performs the best in all cases except for Sumter, for which *Inter* is the best. However, the differences in model performance are magnified for the obstetrics patients.

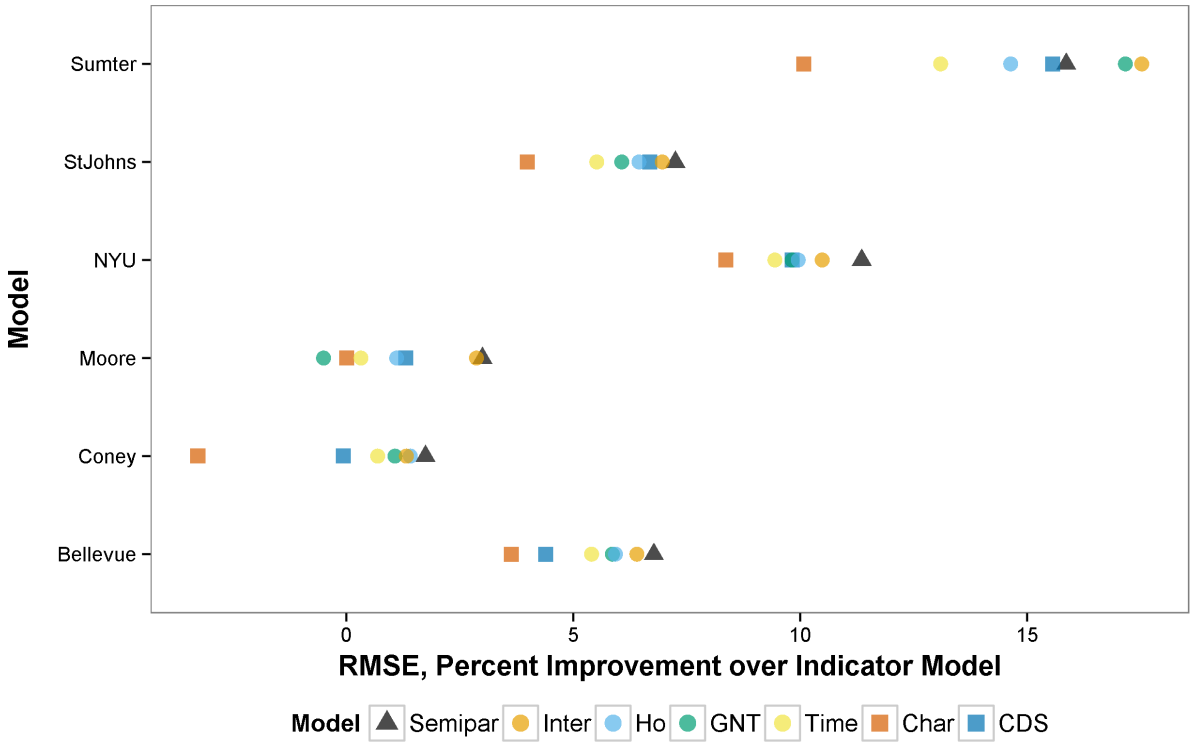


Figure 23 Relative Improvement in Predictive Accuracy at the Individual Level Across Hospitals: Cardiac Sample

Note: Improvement is Percentage Improvement in RMSE for each model over the *Indic* model. Only patients with MDC code 5 are included.

D.7 Different Choice Sets

In our baseline specifications, we set our choice set to include any hospital with a share of 1 percent in any period in our data, excluding the period of the disaster itself. This generates large choice sets; on average, we have between 12 and 21 hospitals in our choice sets. These large choice sets mirror the literature; for example, Capps et al. (2003) has a choice set with 22 hospitals, Ho (2006) reports an average of 15 hospitals per market, and Gowrisankaran et al. (2015) has a choice set of 11 hospitals, with one hospital with a share below one percent. In this section, we examine how model performance changes as we vary the threshold for inclusion from 1 percent to 6 percent, examining each integer value in between. This procedure can vary the choice set quite a lot – for example, St. John's and Sumter have 21 and 15 hospitals given a cutoff of 1 percent, and 4 and 6 hospitals given a cutoff of 6 percent.

Figure 25a and Figure 25b depicts how the model performance of *Semipar*, *Inter*, and *CDS* varies as the choice set changes at the aggregate and individual levels. The x-axis is the number of hospitals in the choice set and the y-axis the relative performance of each model compared to *Indic*. At the aggregate level, *CDS* improves relative to the other two models as the choice set

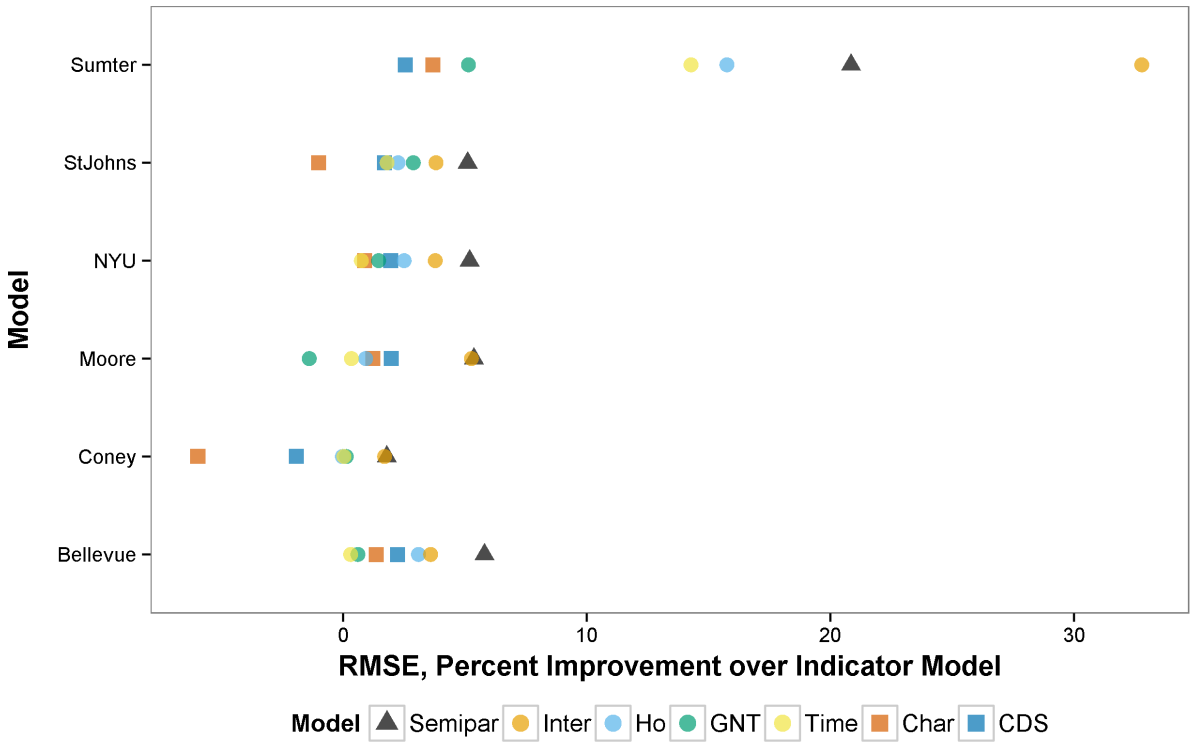
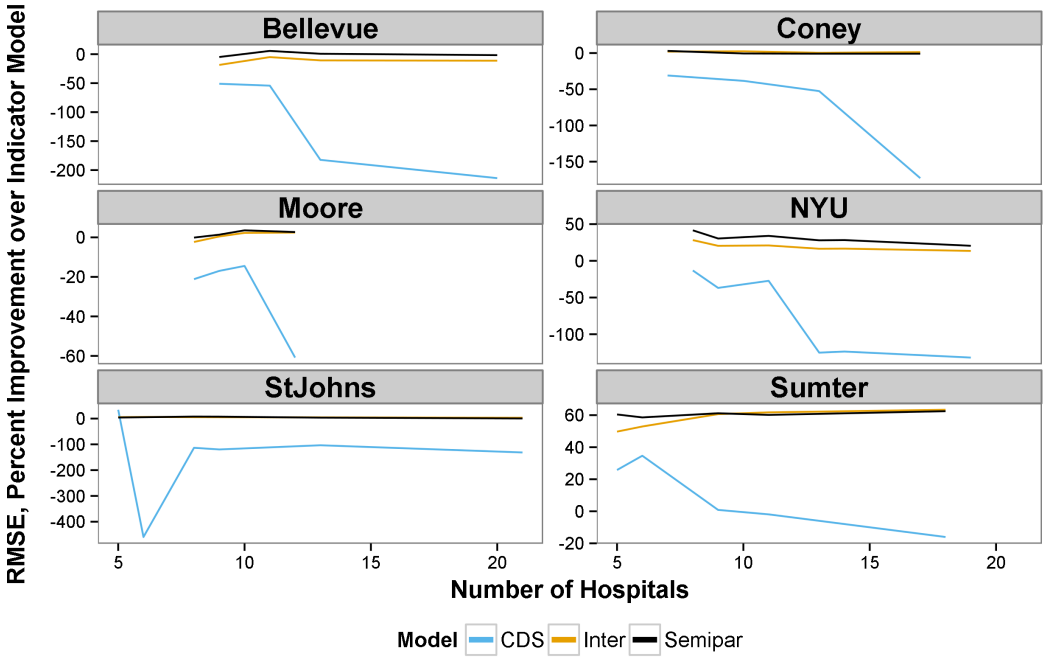
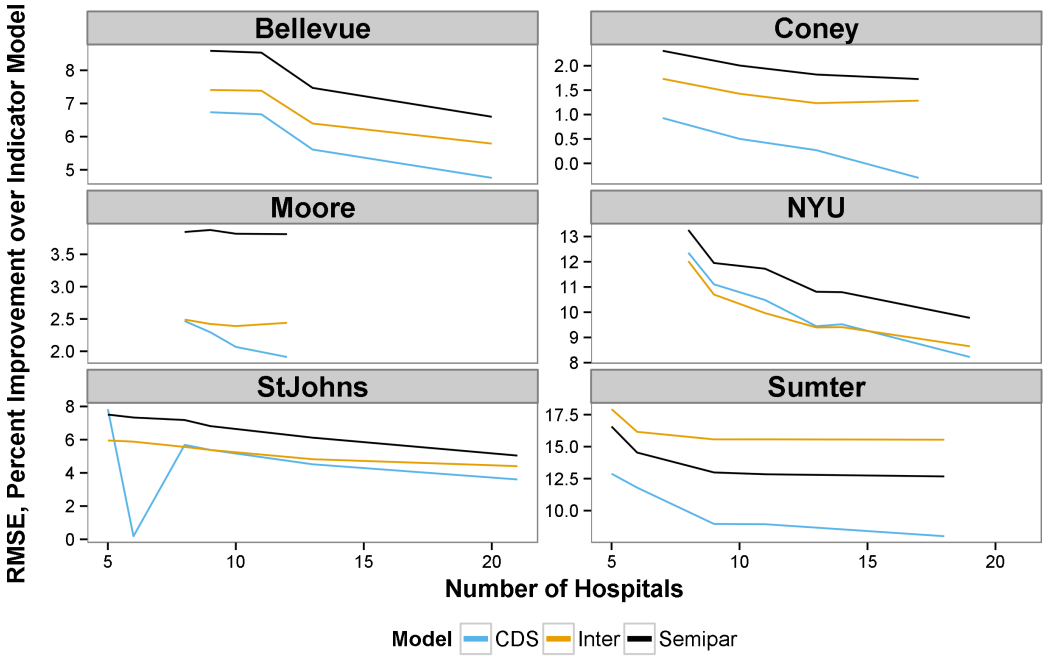


Figure 24 Relative Improvement in Predictive Accuracy at the Individual Level Across Hospitals: Obstetrics Sample

Note: Improvement is Percentage Improvement in RMSE for each model over the *Indic* model. Only patients with MDC code 14 are included.



(a) Aggregate



(b) Individual

Figure 25 Relative Improvement in Predictive Accuracy Across Hospitals as Choice Set Changes

Note: Improvement is Percentage Improvement in RMSE for each model over the *Indic* model. We vary the choice set cutoff from 1 percent to 6 percent, examining each integer value in between.

shrinks, although it is always significantly worse than the other two models. As the choice set shrinks, the model performance of all three models improves at the individual level. Intuitively, the hospitals with a small market share were more difficult to predict. Second, however, *CDS* does relatively better as the choice set shrinks. While *CDS* is the worst model of the three for the largest choice set, it is better than *Inter* for two of the experiments for the smallest choice set. One possible explanation is that it is easier to map the observed characteristics in *CDS* to the unobserved characteristics space when the choice set is small. When the choice set is large, such a mapping may not be possible and so the pure characteristics models perform much worse.

D.8 Different Performance Measures

So far, we have focused on the RMSE. We find similar results across a wide set of goodness of fit measures. In this section, we examine these different measures for individual level prediction.

The first measure is the mean absolute error (MAE), defined as:

$$MAE = \frac{1}{N} \sum_i \sum_j |y_{ij} - \hat{y}_{ij}|$$

The MAE penalizes errors linearly while the RMSE penalizes errors quadratically, so the MAE penalizes large errors much less than the RMSE. The second measure is the relative entropy, or the Kullback-Leibler divergence, of the model, defined as:

$$Entropy = \frac{1}{N_I} \sum_i -log(\hat{y}_{ij^*})$$

where j^* is the alternative actually chosen. The relative entropy is the negative of the average predicted log probability averaged over the actual probability distribution; thus, only predicted probabilities for the choice actually picked are averaged. The relative entropy is zero when the predicted probability for each choice chosen is one, so the model predicts perfectly. Thus, the relative entropy provides a measure of distance between the actual and predicted probability distributions, and even more strongly penalizes larger errors than the RMSE does.

If a model predicts a zero probability for an observed choice, the entropy statistic will be infinite. For the *Semipar* model, this can pose a problem, as the *Semipar* model will give zero probabilities for hospitals that no one within a group went to. We thus bottom code all probabilities by a probability of .001.

For the last measure, we set the predicted choice for each individual as the alternative given the highest probability. This measure, the individual prediction loss, is then the fraction of individuals that we would predict incorrectly:

$$Zero - One\ Loss = \frac{1}{N_I} \sum_i 1(y_i \neq \arg\max_j \hat{y}_{ij})$$

Here y_i is the alternative chosen by individual i. The individual prediction loss would be zero if we predicted each individual's actual choice correctly. This measure is useful in cases for which we would like to provide individuals the most likely prediction. For example, we might want to provide each patient with a default hospital to go to which would correspond closest to their typical choice behavior.

In Figure 26, we plot the relative performance of each model across these four measures at the individual level – the RMSE, MAE, Entropy, and Zero-One Loss – for Sumter. As before, this relative performance is relative to the *Indic* model. The *Inter* model is the best model for the RMSE, Zero-One Loss, and Entropy, while the *Semipar* model is the best for MAE. *Semipar* performs particularly poorly for the Entropy measure because it often predicts zero probabilities that are heavily penalized based upon the bottom code, although it is still better than the characteristics based models. For all other hospitals, the various measures all agree on the best performing model, except that *Semipar* is typically the second best model after *Inter* on the Entropy measure.

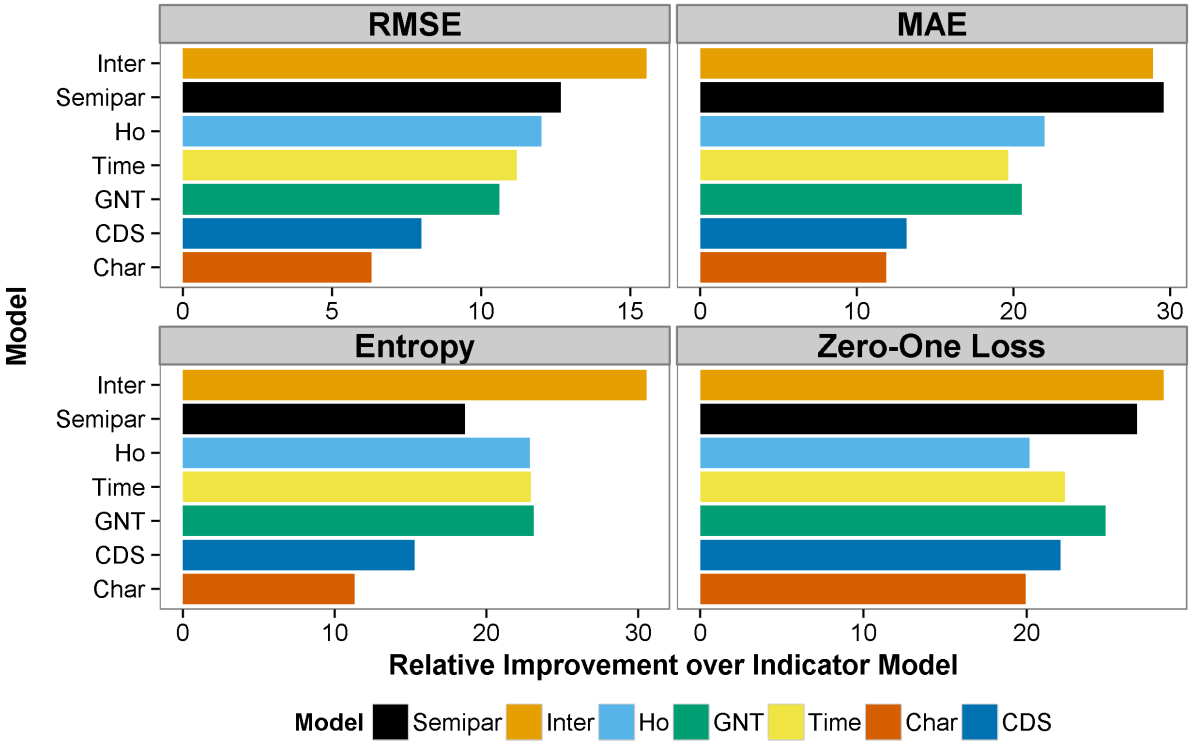


Figure 26 Relative Improvement in Predictive Accuracy at the Individual Level Across Measures for Sumter Hospital Area

Note: Measures are as defined in the text. Improvement is Percentage Improvement for each model over the *Indic* model.

Another way to examine how model performance varies across measures is through the correlations across models of each measure with RMSE. Table D-7 examines these correlations for each

disaster at the individual level. All measures are correlated with RMSE with correlation of at least 0.89; most correlations are close to one.

Table D-7 Correlation with RMSE, Individual Level

Measure	Sumter	NYU	Bellevue	Coney	Moore	StJohns
MAE	0.97	0.99	0.99	0.99	0.97	0.97
Entropy	0.97	0.99	0.99	0.99	0.97	0.98
Zero-One Loss	0.91	0.98	0.89	0.95	0.97	0.94

Note: Each correlation is across models, including *Indic*, for the given hospital service area.

E WTP and Unobserved Heterogeneity

In this section, we formally demonstate how changes in the probability distribution for hospitals across patients would affect the percent change in WTP.

Define $f(s_1, s_2)$ as:

$$f(s_1, s_2) = \log[\frac{1}{1 - s_1 - s_2}] - \log[\frac{1}{1 - s_1}] - \log[\frac{1}{1 - s_2}] \tag{5}$$

This expression is the change in WTP for someone with probabilities s_1 for hospital 1 and s_2 for hospital 2.

Each econometric model produces a distribution of s_1 and s_2 across people. The aggregate change in WTP is just the expectation over f given this probability distribution. The percent change in WTP is the aggregate change in WTP divided by the expectation over $\log[\frac{1}{1-s_1}] + \log[\frac{1}{1-s_2}]$.

We are interested in changes to this probability distribution that increase the share of consumers with high probabilities for both hospitals or low probabilities for each hospital and decrease the share of consumers with high probabilities for one hospital and low probabilities for the other hospital.

Definition 1 *H' is defined to be more concordant than H if, for discrete probability measures H and H', H and H' have the same marginal distributions, and H' can be obtained from H by a finite number of changes that, for (x, y, x', y') such that $x' > x$ and $y' > y$, add mass $\epsilon > 0$ to (x, y) and (x', y') and subtract mass ϵ from (x, y') and (x', y).*

Proposition 1 *Given distribution H' more concordant than H, the expectation of the change in WTP and percent change in WTP is higher under H' than H.*

Proof. By Theorem 1 of Tchen (1980), if H' is more concordant than H, then $H'(x,y) \geq H(x,y)$, where $H(x,y)$ and $H'(x,y)$ are cdfs of the respective distributions. Then, Theorem 2 of Tchen (1980) states that, for any bounded, right-continuous, super-additive function ϕ, and H and H' with identical marginals,

$$\int \phi dH' - \int \phi dH = \int [H'^{-}(x,y) - H^{-}(x,y)]dK \tag{6}$$

where K is the positive measure associated with ϕ, and H^{-} and H'^{-} are the left continuous cdfs of the respective probability measures.

Well, since H' is more concordant than H, $H'^{-}(x,y) - H^{-}(x,y) \geq 0$ for all (x,y), and since K has positive measure the right hand side must also be ≥ 0.

This means that $\int \phi dH' - \int \phi dH \geq 0$.

We now need to show that the WTP function f is bounded, right-continuous, and super-additive. f will be bounded and continuous except where $s_1 + s_2 = 1$; in all practical applications, $s_1 + s_2 < 1$.

We now need to demonstrate super-additivity; i.e., for (x_1, y_1, x_0, y_0) such that $x_0 < x_1$ and $y_0 < y_1$,

$$f(x_1, y_1) + f(x_0, y_0) > f(x_1, y_0) + f(x_0, y_1) \tag{7}$$

We can rewrite this as:

$$f(x_1, y_1) + f(x_0, y_0) - f(x_1, y_0) - f(x_0, y_1) > 0 \tag{8}$$

Now, the LHS is equivalent to:

$$f(x_1, y_1) + f(x_0, y_0) - f(x_1, y_0) - f(x_0, y_1) = \int_{x_0}^{x_1} \int_{y_0}^{y_1} f_{xy}(x,y) dy dx \tag{9}$$

So it should be sufficient to show that the cross-partial f_{xy} is always positive.

$$f_{xy} = \frac{1}{(1-x-y)^2} > 0 \tag{10}$$

Thus, f is superadditive, and we have proved the first part of the claim. The percent change in WTP is the expectation over f divided by the expectation over $\log[\frac{1}{1-s_1}] + \log[\frac{1}{1-s_2}]$. Because the marginal distributions of H and H' are the same, the expectation over $\log[\frac{1}{1-s_1}] + \log[\frac{1}{1-s_2}]$ is the same for both distributions, and so the expected percent change in WTP is also higher under H' than H.

■

F Table of Variables Used

	Indic	Char	CDS	Time	Ho	GNT	Inter
Hospital Indicators	X			X	X	X	X
× Weight						X	X
× Time							X
× Obstetrics							X
× Circulatory							X
× Digest							X
× Muscular							X
× Respiratory							X
× Female Repro							X
Inside		X	X				
× Cardiac Surg Diag			X				
Same County			X				
Time		X	X	X	X	X	X
× Median Income			X			X	
× LOS			X				
× nPX			X				
× nDX			X				
× Emergency			X		X		

	Indic	Char	CDS	Time	Ho	GNT	Inter
× Medical			X				
× Obstetrics			X				X
× Weight		X	X			X	X
× Age						X	
× Under18			X				X
× Over64		X	X			X	X
× Female		X	X			X	X
× Black							X
× Cardiac Surg Diag			X				
× Circulatory			X				X
× Digest			X				X
× Muscular			X				X
× Respiratory			X				X
× Female Repro			X				X
× RN Share			X				
× Teach			X				
× RN Intense			X				
× For Profit						X	
× Beds						X	
× Residents Per Bed						X	

	Indic	Char	CDS	Time	Ho	GNT	Inter
× Teach						X	
Squared Time		X	X	X	X	X	X
× Weight		X					X
× Over64		X					X
× Under18							X
× Female		X					X
× Black							X
× Obstetrics							X
× Circulatory							X
× Digest							X
× Muscular							X
× Respiratory							X
× Female Repro							X
Closest						X	
Cardiac Surg Hosp							
× Cardiac Surg Diag × Adult		X					
× Weight × Adult		X					
Obstetrics Hosp							
× Obstetrics Diag		X					

	Indic	Char	CDS	Time	Ho	GNT	Inter
× Female		X					
NICU Hosp							
× Female		X					
× Obstetrics Diag						X	
Residents/Bed							
× Weight		X					
× Over64		X					
× Female		X					
RN Share							
× Female			X				
× Over64			X				
× Median Income			X				
× LOS			X				
× nPX			X				
× nDX			X				
× Under18			X				
RN Int							
× Female			X				
× Over64			X				

	Indic	Char	CDS	Time	Ho	GNT	Inter
× Median Income			X				
× LOS			X				
× nPX			X				
× nDX			X				
× Under18			X				
RN/Bed							
× Commercial					X		
× Cardiac					X		
× Oncology Alt					X		
× Neurology					X		
× Digest Alt					X		
× Labor and Delivery					X		
× Median Income					X		
For Profit							
× Weight		X					
× Over64		X					
× Female		X					
× Commercial					X		
× Cardiac					X		
× Oncology Alt					X		

	Indic	Char	CDS	Time	Ho	GNT	Inter
× Neurology					X		
× Digest Alt					X		
× Labor and Delivery					X		
× Median Income					X		
Imaging Complexity							
× Commercial					X		
× Cardiac					X		
× Oncology Alt					X		
× Neurology					X		
× Digest Alt					X		
× Labor and Delivery					X		
× Median Income					X		
Teach							
× Female			X				
× Old			X				
× Median Income			X				
× LOS			X				
× nPX			X				
× nDX			X				

	Indic	Char	CDS	Time	Ho	GNT	Inter
× Under18			X				
× Commercial					X		
× Cardiac					X		
× Oncology Alt					X		
× Neurology					X		
× Digest Alt					X		
× Labor and Delivery					X		
× Median Income					X		
Cardiac Complexity							
× Commercial					X		
× Cardiac					X		
× Median Income					X		
× Commercial					X		
× Oncology Alt					X		
× Median Income					X		
Birth Complexity							
× Commercial					X		
× Labor and Delivery					X		
× Median Income					X		

	Indic	Char	CDS	Time	Ho	GNT	Inter
Oncology Diag × Cancer Center			X				
Delivery × Birth Room			X				
Circulatory × Cardiac ICU			X				
Circulatory × Cath Lab						X	
Under18 × Ped Beds			X				
Trauma × CTC			X				
Imaging Diag × MRI						X	

Variable	Source	Description
Adult	Disch	Age greater than 17
Age	Disch	Patient age
Beds	AHA	Number of beds in hospital
Black	Disch	Patient is racially black
Birth Room	AHA	Whether hospital has birthing (LDR, LDRP) room
Birth Complexity	AHA	We apply the Ho services intensity algorithm (see below) to the obstetrics and birth room flags from the AHA data
Cardiac Surg Diag	Disch	For V24 DRG coding: DRGs between 215 and 236; Between V24 and V12: DRGs in this list (104,105,106,108,515,525,535,536,547,548,549,550); Below V12: DRGs between 103 and 107, 207[46]
Cardiac Surg Hosp	AHA	Whether hospital has an cardiac surgery program
Cardiac Diagnosis	Disch	3 digit ICD9 diagnosis codes between (and including) 393 and 398, 401 and 405, 410-417, 420-429

[46]Across the models, three different underlying variables are based on the patient's diagnosis. First, the discharge data include ICD9 diagnosis codes for patients; these diagnosis codes, along with other variables such as procedures, age, sex, discharge status, and the presence of complications or comorbidities, are used to assign a Diagnosis Research Group or DRG. The DRGs themselves are grouped into 25 different Major Diagnosis Categories or MDCs. For example, a patient presenting signs of "maple syrup urine disease" would have ICD9 diagnosis code 270.3, DRG 642 (Inborn and other disorders of metabolism), and MDC 10 (Diseases and Disorders of the Endocrine, Nutritional And Metabolic System).

Cardiac ICU	AHA	Whether hospital has cardiac ICU
Cardiac Complexity	AHA	We apply the Ho services intensity algorithm (see below) to adult diagnostic catheterization, cardiac intensive care, adult interventional cardiac catheterization, and adult cardiac surgery flags from the AHA data
Cath Lab	AHA	Whether a hospital has both a diagnostic and interventional catheterization lab
Cancer Center	AHA	Whether hospital has oncology services
Cancer Complexity	AHA	We apply the Ho services intensity algorithm (see below) to the cancer and maximum of the image-guided radiation and intensity-modulated radiation flags from the AHA data
Closest	Disch	Whether hospital is closest facility to patient
Circulatory	Disch	MDC equals 5
Commercial	Disch	Patient has a commercial insurer
CTC	AHA	Certified Trauma Center
Delivery	Disch	For DRG coding above V24: DRGs in this list (765, 766, 774, 775, 767, 768, 776, 769, 777, 780, 781, or 782). For DRG coding below V12: DRGs from 370-378 and 382-384
Digest	Disch	MDC equals 6
Digest Alt	Disch	3 digit ICD9 diagnosis codes between (inclusive) 520 and 579
Emergency	Disch	Patient admitted through emergency room
Female	Disch	Patient is female
Female Repro	Disch	MDC equals 13
For Profit	AHA	Whether hospital is a for profit facility
Imaging Complexity	AHA	We apply the Ho services intensity algorithm (see below) to SPECT, MRI, CT, ultrasound, and PET scan flags from the AHA survey
Imaging Diag	Disch	MDC code is 1, 5, or 8
Labor and Delivery	Disch	ICD9 diagnosis codes between (inclusive) 650 and 657, 644, 647, 648, V22, V23, V24, V27
Inside	NA	Hospital is not the outside option
Median Income	ACS	Median income of zip code
Medical	Disch	Medical DRG
MRI	AHA	Hospital has an MRI
Muscular	Disch	MDC equals 8
nDX	Disch	Number of diagnoses

nPX	Disch	Number of procedures
LOS	Disch	Length of stay
Neurology	Disch	3 digit ICD9 diagnosis codes between 320 and 326, 330 and 337, or 340 and 359 (inclusive)
NICU Diag	Disch	For V24 DRG Coding: DRG 790 or 791; Pre V24: DRG 386 or 387
NICU Hosp	AHA	Hospital has a NICU
Obstetrics Diag	Disch	MDC equals 14
Obstetrics Hosp	AHA	Hospital has an obstetrics program
Oncology Diag	Disch	MDC equals 17 or for DRG later than V24 in this list (54, 55, 146, 147, 148, 180, 181, 182, 374, 375, 376, 420, 421, 422, 435, 436, 437, 542, 543, 544, 582, 583, 584, 585, 597, 598, 599, 656, 657, 658, 686, 687, 688, 711, 712, 715, 716, 722, 723, 724, 739, 740, 741, 736, 737, 738, 744, 745, 754, 755, 756, 843, 844, 835, 836, 837, 838, or 839). For DRG pre V12 in this list (10, 11, 64, 82, 172, 173, 199, 203, 239, 257, 258, 259, 260, 274, 275, 303, 318, 319, 338, 344, 346, 347, 354, 355, 357, 363, 366, 367, 406, 407, 408, 413, 414).
Oncology Alt	Disch	3 digit ICD9 diagnosis codes between 140 and 239 (inclusive)
Over64	Disch	Patient is over 64 years old
PatCounty	Disch	Patient's County of Residence
Ped Beds	Disch	Hospital has pediatric beds
Respiratory	Disch	MDC equals 4
Residents Per Bed	MCR	Residents per bed from Medicare Cost Reports
RN Share	AHA	Nurses regularly working as a share of licensed nurses
RN Intense	AHA	Nurses regularly working as share of inpatient days
RN/Bed	AHA	Nurses per bed
Same County	Disch	Hospital and patient in same county
Teach	AHA	Teaching hospital
Trauma	Disch	MDC equals 24
Time	Disch/Compute	Travel time from centroid of patients zip code to hospital
Under18	Disch	Patient is under 18
Weight	Disch	DRG weight

Description of Ho Services Intensity Algorithm: Hospitals were rated on a scale of zero to one, reflecting the sophistication of their services in different categories. Zero indicates low sophistication and one indicates a high level of sophistication. The four categories are cardiac,

imaging, cancer, births.

The intensity variable for category c in hospital h is given by:

$$\max\{\max_{x \in X_c} 1_{xh} * (1 - \bar{x})(1 - \bar{y_c}), 1_{yh_c}\}$$

where

- x indexes the services in each category
- 1_{xh} is 1 if hospital h offers service x and 0 if not
- $\bar{x}$ is the state share of hospitals offering that service
- y is the service with the smallest $\bar{x}$
- 1_{yh} is 1 if hospital h offers service y and 0 if not
- $\bar{y}$ is the percent of hospitals offering service y

For more details see Table IX in Ho (2006).