Comparative Advantage in Health Care Delivery: A Machine Learning Approach

Grant Gannaway

Click Here for Most Recent Version

Abstract

With health care spending having increased roughly 35% from 2010 to 2017, now consuming over $3 trillion per year in the US alone, there is growing interest in ways of reducing costs without compromising health outcomes. Since a large share of health care costs come from labor, one approach many states have taken is to change regulations to expand the set of medical providers, shifting from just medical doctors (MDs) to increasingly allow for non-doctors (NDs), such as nurse practitioners, as well. Because ND salaries are so much lower than MDs on average, the hope is to capitalize on their potential comparative advantage in providing routine care to low-risk patients. But there is also the logical possibility that average care quality declines because of the more limited training of NDs relative to MDs, and/or the possibility that ND caseloads wind up including non-routine cases or high-risk patients, which could create health complications and hence increase costs in the longer term. In this paper I study the effects of ND use on costs and patient outcomes using state law changes as a natural experiment, which provides difference-in-difference-type variation. This identification strategy is limited in the aggregate due to weak instrument bias. However, using modern machine learning methods, I am able to narrow in on the subgroup of patients where the first stage is sufficiently strong to produce accurate results in the second stage. These methods are very data intensive, but in health care (and increasingly throughout the social sciences) large enough data is becoming common, allowing researchers to increasingly capitalize on such methods and more effectively estimate heterogeneous treatment effects. I find that the patients who are most likely to be affected by the policy changes have increased rates of both preventable hospitalizations and total medical spending – that is, increased use of NDs on net has adverse effects for the most relevant sample of patients. Estimates for heterogeneous treatment effects in both the first and second stage equations for my instrumental variables analysis helps us understand why: I show that the patients who are predicted to benefit the most from ND care are not the same patients predicted to shift to NDs after the policy changes, suggesting that improved sorting of patients between provider types could fully exploit comparative advantages and result in improved patient outcomes overall.

1 Introduction

Each year over $3 trillion is spent on health care in the US, of which over 40% ($1.32 trillion) is devoted to outpatient care. Though appropriate spending on outpatient care is often an efficient use of health care delivery, there is nonetheless widespread consensus that the value of the care provided is substantially less than the costs, in part because of exceptionally high health care labor costs in the US (**?**). There is widespread interest in reducing health care costs without reducing quality, and, given the high labor costs, one popular proposal is to shift health care tasks from medical doctors (MDs) to non-doctor medical providers (NDs) such as nurse practitioners. Between 2004 and 2015, 11 states relaxed their scope of practice laws, removing restrictions on the medical procedures that NDs could perform by reducing MD supervision requirements. In this paper, I examine the degree to which this shift in medical care affected costs and quality of care.

How this shift will affect costs and quality is the topic of ongoing policy debate and is not immediately clear as a conceptual matter. On the one hand, proponents of these law changes claim that relaxing restrictions on NDs will reduce medical spending without reducing quality of care[1] (**???**). However, opponents (such as the American Medical Association) argue that it is possible that increased ND use will instead lead to *increased* costs and/or *decreased* quality[2] (**?**).

In this paper, I analyze the effects of ND use on patient health and spending outcomes in a large private insurance medical claims database. To do so, I leverage changes in state laws regulating the use of NDs as a natural experiment, which provides difference-in-difference-type variation. This approach is limited due to weak instrument bias - there are tens of millions of types of medical encounters that are

[1] A key supporting point for this claim is that NDs are significantly cheaper to train and employ than MDs - in 2018 the median nurse practitioner salary was roughly half of the median general practice physician salary (see BLS Occupation Employment Statistics May 2018 for generalist physicians and nurse practitioners).

[2] This is possible, for example, if NDs order more tests, refer to specialists more often, and/or mis-diagnose patients more frequently than MDs. Such outcomes may occur as a form of risk aversion by NDs, or if patients or insurers inaccurately predict patient risk or complexity and sort overly-complex patients to NDs.

simply not "at risk" of being shifted from MDs to NDs. To overcome this issue, a reasonable approach might be to rely on institutional knowledge and theory to develop a model of which types of tasks are most likely to be shifted to NDs after the law changes. Such an approach has tradeoffs: a model may provide potentially interesting counterfactual predictions regarding the effects of shifting different sets of tasks to NDs, but requires relatively strong assumptions regarding the task allocation decisions. Furthermore, in this setting, institutional knowledge is limited- there are many types of tasks about which a model would not be able to make informed predictions about the propensity for the task to shift to NDs. Another approach is to rely on recently developed machine learning methods in combination with very rich data in order to make empirical predictions about which types of encounters are likely to shift to NDs.

Similar to the more traditional modeling approach, the machine learning approach has tradeoffs as well. While machine learning does not require strong assumptions or institutional knowledge, it is not as transparent as a traditional economic model and makes predictions on complex interactions of many different observable features of each encounter. Given the ambiguity of theoretical predictions based on institutional knowledge, and the availability of large data, in this paper I opt for the machine learning approach. Doing so yields predicted first stage effects - the predicted change in the probability of ND use - for each encounter in the data. I use these predicted effects to remove tens of millions of encounters that have low probability of shifting to NDs, and focus instead on types of encounters with high first stage signal. This is similar in spirit to the approach used in **?**. These methods are data intensive, but the large health insurance claims data I use allows for such an approach. Increasingly, large enough data is becoming more widespread in the social sciences so that researchers will be able to capitalize on the benefits of such data intensive methods.

Interestingly, conclusions based on this high first stage sample are significantly different from those based on the overall sample. In the overall sample, I find an exogenous increase in ND use, but no significant changes in preventable hospitalizations or spending. In the set of patients most likely to be shifted to NDs on the

other hand, I find statistically significant increases in 1-year preventable hospitalization rates. Instrumental variables (IV) estimates suggest NDs cause a roughly 0.5 percentage point increase in 1-year preventable hospitalizations in this group, up from a baseline level of 0.2 percent of encounters that were followed by a preventable hospitalization prior to the law changes. The effect on total spending among these patients is imprecise, with a positive point estimate.

To understand the mechanisms behind these results, I estimate the reduced form effects of the law change on potential drivers of the increased hospitalization. I find that the high first stage sample has a decrease in medication adherence and no increases in prescription fill rates or "diversity" of prescription types[3], relative to the overall sample, which had no decrease in medication adherence and increases in prescription fill rates and diversity. Furthermore, the overall sample saw increases in outpatient and prescription spending, while the high first stage sample instead had increases in inpatient spending.

I seek to understand whether the observed increases in the outcome variables could be mitigated had the sorting of patients between provider types been adjusted. To study this question, I utilize the predictions on the effect of shifting to an ND (mentioned above) on the *full sample*, instead of just the high first stage group. I combine these predictions with encounter-specific machine learning predictions of the *IV effect* of ND use on patient outcomes. This is analagous to comparing heterogeneous first stage estimates to IV estimates in a traditional analysis. The benefit of this exercise is that I estimate, *for each type of encounter*, the probability of being shifted to NDs as well as the predicted changes in outcomes from ND use. Comparing the two predictions gives an estimate for the efficiency of the sorting of encounters between provider types and provides an answer to the question about whether the types of encounters that *are* shifted to NDs are the same types of encounters that *should be* shifted to NDs.

The sorting analysis reveals that the sorting of encounters between provider types is less than ideal. Considering only the predicted change in spending from ND use,

[3]Prescription diversity here refers to the number of unique types of prescription therapeutic classes a patient uses over the following year.

actual sorting is almost exactly the opposite of a sorting aimed solely at reducing spending. There are relatively few types of encounters that are both strongly predicted to shift to NDs *and* predicted to reduce spending as a result of ND use. Instead, there are many types of encounters that are *not* predicted to shift to NDs but *are* predicted to *reduce* spending as a result of ND use. When considering the effect of ND use on preventable hospitalizations, the results still leave room for improvement: I find a large cluster of types of encounters that are predicted *not* to be shifted to NDs and are also strongly predicted to *increase* hospitalizations as a result of ND use. However, the largest cluster of types of encounters are predicted to decrease hospitalizations as a result of ND use, but are only *moderately* likely to be shifted to NDs. As with spending, there are relatively few types of encounters that are strongly predicted to shift to NDs as well as decrease hospitalizations as a result of ND use.

Taking the results together suggests that even though relaxing restrictions on NDs does not show adverse outcomes in the *overall* set of encounters, this aggregate effect masks substantial heterogeneity. When the heterogeneity is uncovered and exploited, NDs are shown to statistically significantly increase patient preventable hospitalizations on the group of patients most affected by the law changes (with imprecise estimates of increased spending). These effects appear to be driven by the types of encounters sorted to NDs after the law changes, as well as the ND effects on prescription drugs and the substitution of inpatient for outpatient care. By improving sorting of encounters between provider types, comparative advantage could be fully exploited and patient outcomes could be improved.

This paper contributes to the larger health economics literature that tries to understand ways of reducing low-value care without compromising patient health, in this particular case by allocating tasks between provider types to better take advantage of provider specialization. Other papers have studied the effects of state scope of practice law changes and ND use, but my application of modern statistical methods allows me to answer this question beyond the current literature since I am able to narrow in on the most relevant set of patients as defined by an extensive multi-dimensional set of observable factors. I also contribute to the machine learning

for heterogeneous treatment effects literature by providing a blueprint for applying recently developed machine learning methods for estimating heterogeneous treatment effects in randomized experiments to natural experimental settings, and in cases with small observed first stage effects.

The rest of the paper proceeds as follows. In section 2 I first explain the background of the outpatient medical setting and the scope of practice laws. Then I explain my contribution to the literature on relaxing scope of practice laws and the effects of NDs, as well as to the literature on machine learning for heterogeneous treatment effects. In section 3, I give details on the claims data I use in this paper. Section 4 explains the traditional econometric methods and gives their results in the overall sample. In section 5, I outline the machine learning methods I use, and in section 6 I show the results from these methods in the high first stage sample. Section 7 concludes.

2 Background Setting

NDs are essentially nurses with master's degrees. The master's degrees are usually received at medical schools after completing a two or three year program. The programs do not attempt to cover everything taught in medical school, but instead focus on mastering a subset of medical skills. To practice as an ND, candidates must obtain licenses from the state in which they intend to practice, and must abide by state scope of practice (SOP) laws.

State SOP laws are generally determined by state medical licensing boards in conjunction with state legislatures. These laws primarily regulate the level of physician supervision required for ND practice. There is substantial heterogeneity in the restrictiveness of state laws, with some states allowing NDs to both provide medical treatment and prescribe medication without physician supervision, while other states require NDs to obtain physician approval for both treatment and prescription decisions. A primary mechanism for enforcement is malpractice litigation: in restrictive states, MDs bear the primary responsibility for malpractice. Thus, insurance companies are less likely to cover ND care in restrictive states.

There has been a recent trend toward relaxing SOP laws, with at least 15 states making some changes to their SOP laws since 2000. The specific timing of changes for each state are shown in table 1, which I compiled based on records of state legislatures and state medical boards. In 2002, the Federal Trade Commission held hearings on the level of competition between different medical provider types in the outpatient setting, which likely pressured states into relaxing laws. The major barrier to relaxing SOP laws are powerful state lobbying groups working to protect physician monopoly power, combined with substantial physician representation on state medical boards.

Facilities vary widely in their approach to sorting of tasks between provider types. Large academic hospitals may never sort any tasks to NDs alone, while small rural practices or clinics may rely on NDs for a majority of tasks. In this paper I focus only on outpatient settings. Since MDs can generally bill at a higher rate, if an ND *can* bill at the MD rate, I assume they do. This biases my results in the sense that any observed ND encounter is strictly an ND encounter, while MD encounters may be a mix of MDs and NDs. This makes the MD effects look similar to the ND effects, biasing any differences I find in the reduced form outcome effects toward zero.

For the first stage effects, I argue that the observed increase in the share of ND encounters is driven by an actual change in which provider type provides the care, and not just a change in billing labels. I support this argument with two main points. First, I observe a reduced form outcome effect at the time of the SOP law changes in the high first stage sample. If care practice remained unchanged and billing labels only changed, there should have been no change in any outcome variables at the time of the SOP law change. However, if care practice *did* change, the changes in outcome variables are justified. Second, the incentive for NDs to bill as "incident to" MDs remains consistent before and after the law change, but MDs are no longer required to supervise ND care. So even after the law change, NDs still have an incentive to bill at the MD rate whenever possible.

Changes in SOP laws might affect the NP share of encounters via multiplie potential mechanisms. As discussed in **?**, a primary burden of SOP restrictions is the administrative burden they create. When SOP laws are relaxed, both MDs and NDs

can reduce their time spent on administrative tasks and increase time spent on care for patients. On the extenisive margin, NDs may be induced to transfer to markets with less restrictive SOP laws (as discussed in **?**).

2.1 Literature

Several recent economics papers have explored the effects of SOP laws on various aspects of the health care industry. Most similar to this paper are **?**, **?**, and **?**. **?** show that SOP law relaxations lead to higher wages for nurse practitioners but lower wages for general practice doctors. They also show that there is little change in the price of a specific procedure that is likely to be performed by both nurse practitioners and general practice doctors (well-child visits). Instead of focusing on provider wages and medical prices, I focus on patient outcomes.

? uses in-depth survey data to show that nurse practitioner independence increases the frequency of routine checkups, improves perceived care quality, and decreases emergency room visits. They also show that the mechanism behind these effects is based on the reduction of administrative costs and the increase in patient access to care. My paper differs from theirs in that I use claims data instead of survey data, and I focus on the sorting of tasks and specialization of provider types to provide heterogeneous treatment effect estimates at a more granular level. Their survey data is of a wider range of patient types than I use - my data is only privately insured patients who are relatively well employed, while their data includes patients from all insurance types. Furthermore, my paper focuses on different patient outcomes: I focus on hospitalizations and spending, while they focus on access and emergency room visits.

The paper most similar to mine in spirit is **?**, which uses the same claims data I use in addition to Medicare claims data to examine the effects of the SOP law changes on various patient outcomes. They use a slightly different time frame than I use, and thus have different state policy changes driving their variation. Their aggregate findings are consistent with my results in the priavately insured population, though they find improvements in outcomes for the Medicare population. They

use a patient movers design in addition to a traditional difference-in-differences. They do not find any effects on patient access to care or office visit prices. The main differentiation between our papers is that I emphasize the machine learning for heterogeneous treatment effects, allowing me to narrow in on the set of encounters most likely to be affected by the law changes. Another key difference is that I use an IV approach, focusing on encounter-level effects, while they report aggregate effects from the difference-in-differences.

? shows that expanded NP and PA supply has had minimal impact on the office-based healthcare market overall, but utilization has been modestly more responsive to supply increases in states permitting greater autonomy. **?** studies another important and related set of providers who are impacted by SOP laws: nurse midwives. They find that states with relaxed restrictions on nurse midwives have lower probabilities of C-section deliveries, and improved birth outcomes.

In the retail clinic setting, **?** study costs created by SOP laws, concluding that up-front costs are higher in retail clinics because of restrictive SOP laws. Other papers have studied the differences between NDs and MDs in a variety of settings. **?** study both the intensive and the extensive margins of patient care to show that retail clinics serve to increase medical care utilization and thus total spending. **?** use a small sample survey to test whether nurse practitioners provide less accurate diagnoses than physicians in the emergency care setting. They find that there are small differences in only a few categories.

The main contributions of my paper to the occupational licensing in health care literature are that I use claims data to measure the effects of care from NDs at the encounter level, using patient outcomes hospitalizations and spending. The application of the machine learning methods to this granular data allows me to estimate *encounter specific* effects, whereas other papers measure only aggregate impacts. Thus, I provide estimates suggesting which groups of patients benefit from ND care, and which don't, instead of the aggregrate effect of ND use in general. Furthermore, the claims data I use also allows me to study the effects of ND use on the privately insured population, a population responsible for a significantly large share of total medical spending.

I also contribute to the literature on machine learning for heterogeneous treatment effects. I use a generalized random forest (GRF) approach which was developed in **?**, **?** and **?**. Other papers have used this approach in different setttings. For example, **?** show heterogeneous effects of summer work programs on different types of young adults. I contribute to this literature by providing an example of adapting the method to a difference-in-difference setting. The difference-in-difference framework requires some slight modifcations to the traditional approach. Before implementing the GRF, I first residualize the data, removing the variation from state and year fixed effects as well as other variables with variation at levels higher than the encounter level from each variable. This residualization is similar to one of the methods compared in **?**, which showed that the residualized GRF did well compared to other HTE methods in non-experimental settings.

3 Data

The primary data source for this paper is the Truven MarketScan claims database, a collection of insurance claims from private insurance agencies and businesses that provide health insurance for their enrollees/employees. Insurance agencies and businesses submit claims for their enrollees to Truven, where the claims are anonymized and aggregated across submitters[4]. Submitters of claims can then access the aggregated data and analyze market-wide trends in an effort to improve their insurance coverage. I use data from years 2003 to 2017.

I use MarketScan data containing information on patient outpatient visits, hospital admissions and stays, prescription drug fills, and emergency room visits. Variables available include payment information, diagnostic information, procedure information, patient home state, and provider type codes. Patients are uniquely identified by an identification number that is consistent across years and types of claims. Substate patient geographic information is not available for my entire sample period, so

[4]Claims are not collected randomly, but rather as a convenience sample of submitting agencies. The non-randomness of the sample does not present econometric identification issues as long as the decision to submit claims is not correlated with changes in state SOP laws, which seems plausible.

I do not include it.

The primary outcome variables I use focus on the year following each outpatient visit. First, I use an indicator for whether the outpatient visit was followed by a "preventable" hospitalization within the next year. To define "preventable" I use the standard definition[5] from the Agency for Healthcare Research and Quality which lists the diagnostic codes for 13 prevention quality indicators for diagnoses that are widely considered preventable in an outpatient setting. This variable is a metric of quality of outpatient care. I also show results adjusting this variable to whether the outpatient encounter was followed by a preventable hospitalization within 90 days, and for all hospitalizations (including non-preventable).

The other outcome variable I use is the log of total medical spending by or in behalf of each patient over the year following the outpatient visit, excluding cash payments and premiums. This measure includes copayments, deductible payments, and payments by the insurance company for outpatient, inpatient, and prescription drug claims. Ideally, outpatient visits will be effective in reducing future spending. The combination of this spending variable and the preventable hospitalization indicator reflects health outcomes quality and spending.

I provide details of the procedure I use to obtain the estimation sample in appendix A.

I obtained the law change dates from state legislation and statute records, as well as by contacting state boards of nursing. Table 1 classifies each state into one of three categories based on SOP law change dates. The first category of states are those that changed their law during the data time frame. The second group of states are those which have ND independence, but relaxed their law prior to the start of the data. The final group are those that had not relaxed their law prior to the end of the data time frame. For the states that changed during the data time frame, table 1 lists the law change year. I limit only to these law-changing states. I focus only on practice authority independence, rather than prescription authority since I am interested in the effect of patients receiving care from an ND, which is directly connected to the NP practice authority.

[5] `https://www.qualityindicators.ahrq.gov/Modules/PQI_TechSpec_ICD10_v2018.aspx`

Since the timing of state law changes varies substantially across the data, and the effect of the law change likely takes several years to reach its full effect, I use an unbalanced panel of states which includes all states that changed their laws in the data time frame.

Tables 2 - 3 show summary statistics for the overall sample as well as the high first stage sample. The primary conclusion from these tables is that the high first stage sample is a set of less risky patients who spend less and use the health care system less than the overall sample. The mean age in the overall sample is roughly 35 years old, roughly 40 percent of encounters are male patients, and 13 percent of encounters are from patients from rural (non-MSA) residences. Payments for each encounter and total spending measures have large standard deviations, suggesting a wide range of health care utilitzation and/or intensity of utilization. Roughly 6 percent of encounters were followed by some hospitalization over the year following the encounter, while around 0.6 percent of encounters were followed by a preventable hospitalization over the next year.

4 Overall Sample Approach and Results

4.1 Overall Sample Approach

The primary research question for this paper is whether receiving care from an ND has worse patient outcomes than receiving care from an MD. I exploit changes in SOP laws as a natural experiment that exogenously shifts encounters from MDs to NDs. Thus, the first stage is to measure the effect of the law change on ND utilization. I estimate the first stage effect using a difference-in-differences specification with state and year fixed effects:

$$ND_{ist} = \beta Post_{st} + \gamma_s + \lambda_t + \Theta X_{ist} + \epsilon_{ist} \tag{1}$$

Where γ_s and λ_t represent full sets of state and year fixed effects respectively,

and X_{ist} is a vector of control variables. ND_{ist} is an indicator for whether the encounter was handled by an ND, and $Post_{st}$ is an indicator variable equal to 1 if the encounter occurred in state s after the state relaxed its SOP laws, and 0 otherwise. The coefficient of interest is β, which gives the effect of the law change on the probability of ND use. All standard errors are clustered at the state level. The control variables include encounter-level covariates such as patient age, gender, and rural status (defined as whether the patient lives in an MSA or non-MSA area).

The Affordable Care Act (ACA) contains provisions aimed at increasing the use of NDs. Some states relaxed their SOP laws after the ACA was passed, possibly in response to the federal shift towards increased ND use. These states are plausibly different in the reasons for changing their SOP laws than states that changed prior to the passing of the ACA, and likely have different trends in ND use leading up to the SOP law change. To account for these differences, I include a time trend specific to states that changed their SOP laws after the ACA was passed in each of the specifications.

Given that many state legislatures announced the law changes prior to the changes actually taking place, insurance companies and providers were able to pre-emptively adjust their behavior prior to the policy change. For this reason, I omit encounters from the year prior to the law change in all specifications. This allows me to compare encounters that were "safely" in the pre-law change period to those in the post-law change period.

Since I limit the estimating sample to only states that changed their laws, the primary assumption for causal inference in this case is that the timing of the law changes is conditionally independent of the outcome variable. I verify this assumption by replacing the $Post_{st}$ indicator with a set of years-before/since the law change indicators, and testing whether the coefficients in the pre-treatment period are non-zero. In doing so, I cap encounters that occurred more than six years before or after the law change to six years before/after, which allows me to keep the long run observations in the estimatation sample helping to pin down the treatment effect estimates.

The reduced form effect of the law change on patient outcomes is identical to

equation 1 above, but replaces the outcome variable with a health outcome of interest. The health outcomes I use are an indicator variable for whether the outpatient encounter was followed by a hospitalization for a preventable condition within the next year, and the patient's log total health spending in the next year. I also show results for preventable hospitalizations within the next 90 days and hospitalizations for any reason over the next year. In this reduced form specification, the coefficient on the $Post_{st}$ indicator gives the effect of the law change on the health outcome variable.

To infer causation from the law change on outcome variables, I again take an event study approach as above and show the pre-treatment trends on the years-since indicators with the health outcomes as the outcome variables. If the pre-treatment trend coefficients are not significantly different from zero, this suggests that the law change was conditionally exogenous of the outcome variables of interest.

To measure the effect of receiving care from an ND for those encounters that were moved by the law change, I use an IV approach where the ND indicator is instrumented by the law change. I include the same controls as in the previous specifications. The outcome variables are the health outcomes of interest. The coefficient on the instrumented ND indicator gives the effect of receiving care from an ND for those types of encounters that were moved by the law change from MDs to NDs.

The causal interpretation of the IV results relies on the assumption that the law change did not affect the outcome variables through channels other than the use of NDs. This is plausible given the timing of the law change is as good as random across states (as verified by the event studies on ND use and patient outcomes).

4.2 Overall Sample Results

The results from the traditional first stage estimation on the full sample are shown in the first row of table 4, labeled "NP Usage". The first column shows the baseline (pre-law change) means and standard deviations. The second column (labeled "OLS") shows the effect of the law change (the coefficient on the post law change indicator

in the difference-in-difference regression). As seen in the OLS column, relaxing the SOP laws increases the probability that an encounter is handled by an ND by 3.2 percentage points. This estimate is statistically significant and robust to including encounter-level controls. This effect is up from a baseline share of 2.4 percent of encounters handled by NDs in the pre-law change period.

Table 4 also shows the reduced form outcome effects for the full sample in column 2. None of the outcome estimates are statistically significant at conventional levels. The point estimates suggest a potential decrease in hospitalizations, though relatively large standard errors make any conclusions about these estimates only suggestive at best. In the baseline mean column for the log spending outcome, the entry shows the baseline mean and standard deviation of raw total spending.

Figure 1 shows the effects of the event study regression for the overall sample in blue (and the high first stage subsample in black). The vertical axis gives the estimated coefficient on the years-since indicators, and the horizontal axis shows the years since the law change. The shaded area represents 95-percent confidence intervals based on standard errors clustered at the state level. The years since coefficients in the pre-treatment period are not statistically different from zero, and the point estimates show no trend prior to the law change. After the law change, the coefficients increase steadily up to roughly 13 percentage points six years after the law change. As mentioned above, the encounters in the year prior to the law change are omitted to avoid contamination from anticipation effects.

The dynamic reduced form effects on health outcomes in the full sample are shown in figures 2-4. As with the first stage estimates, the full sample coefficients are shown in blue. None of the outcomes show significant trends in coefficient estimates either before or after the law changes, though the standard errors do not rule out decreases in preventable hospitalizations over the next year.

The third column of table 4 (labeled "IV") shows the results from the traditional IV specifications in the full sample. The coefficients on each type of hospitalizations are negative (though again with relatively large standard errors). The IV coefficient on spending is large and positive, but again, the standard error is large. Interpretted as a local average treatment effect, these estimates do not rule out decreases in

preventable hospitalization for those encounters that were shifted to NDs as a result of the law changes.

Table 4 reports the first stage F-statistic for the IV specification at roughly 7.3, which suggests the identification suffers from a weak instrument problem, and motivates alternative approaches.

As expected, relaxing SOP laws exogenously increases the utilization of NDs, though the changes do not appear to influence a large number of encounters (relative to the number of encounters handled by MDs). The small share of affected encounters is likely the reason for the imprecise reduced form estimates. That is, there are many encounters that have no probability of being shifted to an ND which cloud the identification of the true effect.

Narrowing in on the set of encounters most likely to be impacted by the law change will allow me to increase the power of the first stage estimates. I will then estimate the true effects of receiving care from an ND using the high-first stage set of encounters. To objectively find the high-first stage set without arbitrarily manipulating the data sampling, I use machine learning to predict the first stage effect on the probability of ND use for each type of encounter, and then partition the data into groups based on these predictions. I explain the details of this approach in the next section.

5 Heterogeneous Treatment Effect Methods

5.1 Generalized Random Forest (GRF)

In this section, I briefly explain the details of the machine learning method I use to estimate heterogeneous treatment effects (HTEs). The specific algorithm I use is the generalized random forest (GRF), and a more detailed explanation of the algorithm is found in **?**. I also explain the details of the method in my application in appendix B.

At a high level, the GRF essentially estimates a different weighted least squares regression for each encounter in the data. All encounters are used in each regression,

but a different set of weights is used for each encounter. The weights for a specific encounter, encounter i, are defined so as to place higher weight on encounters that have "similar" predicted probabilities of shifting to NDs as encounter i. "Similarity" is determined using a recursive partitioning approach: encounters are partitioned into subgroups based on cutoffs of predictor variables (ie. age less than 40). Cutoffs are decided by examining the two different probabilities for shifting to NDs in the two resulting subgroups, and then choosing the cutoff that maximizes heterogeneity in these two estimates. This process continues recursively until the partitions reach a minimum number of encounters. Then, encounters in the same partition as i get higher weight than encounters in separate partitions.

So, for example if the first proposed cutoff is age less than 40, the algorithm estimates the probability of shifting to NDs in the age less than 40 subgroup as well as the age greater than or equal to 40 subgroup. The variance of these two estimates is compared to the variance of all other potential cutoffs (such as the male vs female split). The propsed cutoff with the largest resulting variance is executed, and the process repeats in each of the two subgroups. Subgroups are divided until a resulting subgroup reaches a minimum specified number of encounters. Then, all encounters in the same final subgroup as encounter i get higher weights than encounters in different subgroups.

The GRF is designed for experimental settings with exogenous treatment, so I adapt the basic GRF for my quasi-experimental difference-in-difference setting by residualizing out the variation from variables that vary at levels greater than the encounter level from the outcome and treatment variables prior to implementing the algorithm[6]. Thus, I pass the residuals from the outcome and post-law change variables into the algorithm along with the set of predictor variables. The specific variables I residualize out are the state and year fixed effects, and the linear time trend that is specific to the states that changed their SOP law after the passing of the ACA.

The results of the GRF are encounter-specific estimates for the effect of the law

[6]This is similar to one of the methods for estimating HTEs in **?**, which found that the adapted GRF performed well relative to other methods

change on ND usage, which I then collapse to the patient level by taking the median of all predictions for each patient. I then divide patients into deciles based on their median predicted first stage probability of shifting to NDs. I then estimate seperate event study regressions, identical to the first stage and reduced form event studies above, in each decile. Thus, I obtain an estimate for the effect of the law change on ND use, hospitalizations, and total spending in each decile of predicted first stage effect.

After estimating the first stage and reduced form effects for each decile, I then estimate an IV specification for each decile using the same specification as above, where the law change indicator serves as an instrument for the ND-use indicator. Doing so gives an IV estimate for the effect of receiving care from an ND for those encounters moved by the law change, separately for each decile of predicted first stage effect. I focus on the top decile of patient median predicted first stage effect, which is the set of patients most likely to be shifted to NDs after the law change. I show that in this group, the law change is no longer a weak instrument, sharpening the analysis.

5.2 Instrumental Variables Generalized Random Forest

The final questions I turn to are focused on the efficiency of the sorting that results from the law change. That is, I seek to understand whether the types of encounters that *are* being sorted to NDs are in fact the types of encounters that *should be* sorted, in view of reducing hospitalizations and spending. There are other reasons for sorting that I do not capture with these two variables, but these two are important aspects of the health system and of the patient's health care experience, and are widely used measures of quality of care.

I run another machine learning algorithm that is very related to the GRF - the instrumental variables generalized random forest or IV GRF. The process for estimating this algorithm is identical to the GRF, but instead of estimating an encounter-specific first stage effect, the IV GRF produces an estimate for the *IV effect* for each type of encounter. Thus the result of the IV GRF is an encounter-specific prediction

for the effect of care from an ND.

I interpret the IV GRF estimates as the quality differential for each encounter between an ND and an MD. If the estimate is negative, that suggests shifting the given type of encounter to NDs would result in *decreases* in spending and hospitalizations, relative to not being shifted to NDs.

To examine the sorting efficiency, I show a heatmap with the decile of predicted first stage GRF shifting probability effects on the x-axis and decile of predicted IV effects on the y-axis. If sorting were perfectly efficient, the types of encounters for which outcomes are predicted to improve after shifting to NDs should be the most likely to be sorted to NDs after the law change. This would be represented by a concentration of encounters in the bottom right quadrant of the IV GRF-GRF heatmap: most likely to be shifted to NDs (furthest to the right) and largest predicted reductions in spending and hospitalizations resulting from ND care (furthest to the bottom).

6 Heterogeneous Treatment Effect Results

Tables 3 shows the means and standard deviations for the highest first stage decile. As mentioned above, this table shows that the high first stage sample is a younger, less risky group that spends less on medical care and uses the system less frequently than the overall sample. The high first stage group spends less overall, and per encounter, and are hospitalized less frequently than the overall sample. This suggests the patients most likely to shift to NDs are the lower complexity patients, a fact consistent with predictions that would be made using institutional knowledge.

The distribution of predicted first stage effects is shown in figure 6. To assess the accuracy of the predicted first stage effects, figure 7 shows the traditional first stage effects estimated in each decile of predicted first stage effects. The results show a monotonic increase in traditional effect estimates as predicted effects decile increases. In the top decile of predicted effect, the traditional DD estimate is roughly 0.2, implying that in this group, ND use increased by 20 percentage points after the law changes. This top decile is the high first stage sample I use as the relevant sample

of patients affected by the law change.

Figures 1-4 show the event study results for the high first stage sample in black, as well as the overall sample in blue, where the high first stage sample is defined as the top decile of person median predicted first stage effects. The horizontal dashed lines show the DD regression estimates. Figure 1 shows the primary purpose of the GRF: the first stage estimates in the high first stage sample are considerably larger than the overall estimates. In the pre-law change period, the high first stage sample has more negative coefficients than the overall sample, but none of the coefficients are statistically significantly different from 0. This illustrates a potential shortcoming of the GRF: it is designed to find heterogeneity in the summarized DD estimates, regardless of pre-trends (though that does not appear to be an issue in this case).

The summarized DD estimates for the high first stage sample are shown in table 5. The first stage effect is much larger in the high first stage sample than in the overall sample; roughly a 20 percentage point increase in the share of ND encounters after the law change, though the baseline mean is also higher than in the overall sample. The number of encounters in the high first stage sample is not equal to ten percent of the overall number of encounters since the high first stage sample is determined by the patient median predicted effect. Ten percent of patients are in the high first stage sample, but these patients do not have as many encounters as patients in other deciles.

Column 3 of table 5 shows the results from the IV specification in the high first stage sample. First, it is important to note that the first stage F sample is higher in the high first stage sample than in the overall sample, as indicated at the bottom of the table. All IV coefficients in the high first stage sample are positive, and the estimates for the preventable hospitalization outcomes are statistically significant. These results suggest that NDs increase preventable hospitalization rates by roughly half of a percentage point for those encounters shifted to NDs after the law change.

Interestingly, the reduced form outcome effects in the high first stage sample are generally positive (as opposed to negative in the overall sample). This is also reflected in the high first stage sample event studies in figures 2-4. Preventable hospitalization rates over the next year increased by roughly one-tenth of a percentage point, up from

a baseline level of 0.2 percent. Preventable hospitalizations in the next 90 days and hospitalizations for any reason over the year following are not significantly different from 0, but are both positive while their full sample counterparts are negative. Log spending does not change in a significantly different amount in the high first stage sample from the overall sample.

To understand the mechanisms behind these results, I estimate the reduced form DD effects of various other outcome variables in the top decile of predicted first stage effects and in the overall sample. Figure 8 shows the results of these regressions for both the high first stage sample and the overall sample for outcomes related to prescription drugs. This figure shows an increase in prescription fills for both chronic and acute drugs the overall sample, while the high first stage sample does not. The high first stage sample instead has a decrease in mean medical possession ratio (a measure of medication adherence), while the overall sample does not. The overall sample sees an increase in the number of unique types of prescription drug therapeutic classes while the high first stage sample does not. This suggests a potential mechanism behind the observed increase in preventbale hospitalizations: changes in prescription drug behavior. This also may explain why the effect is observed at the one-year level and not the 90-day level, since the effects of not taking prescription drugs may take more time to manifest themselves.

The results of the IV GRF are shown in figures 9 - 10, in comparison to the predicted first stage GRF results. Figure 9 shows a heatmap of the count of encounters in each IV GRF decile - GRF decile bin, where the outcome in the IV GRF is log total spending. Lighter colors represent a higher concentration of encounters. Red lines show deciles above which all predicted effects are positive. Thus, below the horizontal red line represents encounters that are predicted to decrease spending from ND use, while those to the right of the vertical red line are encounters that are predicted to shift to NDs after the law changes. The figure shows largely the opposite relationship between predicted change in spending and predicted probability of shifting to NDs than a sorting aimed solely at reducing spending. The largest cluster of encounters is in the in the middle portion of the figure near the top, which suggests no changes in spending and moderate increases in the probability of ND use.

The next largest cluster of encounters is in the bottom left portion of the figure: low predicted probability of shifting to NDs but large predicted decreases in spending from ND use. There are relatively few encounters in the bottom right of the figure - encounters with high predicted probabilities of shifting to NDs and large predicted decreases in spending as a result of ND use.

Figure 10 shows the sorting of enocunters when the IV GRF outcome is preventable hospitalizations over the next year. This sorting is more encouraging than the sorting with respect to predicted spending, but there is still room for improvement. There is a large concentration of encounters with large predicted decreases in hospitalization rates but only moderate predicted probability of shifting to NDs (the bottom-middle of the figure). There is also a large concentration of encounters in the top left of the figure: predicted increases in hospitalization rates after ND use and low predicted probability of shifting to NDs. There are however, also a relatively large set of encounters in the bottom left of the figure, with low predicted probability of shifting to NDs but large predicted decreases in preventable hospitalizations. There are relatively few encounters in the bottom right of the figure with large predicted decreases in hospitalizations and large predicted probability of shifting to NDs.

An important caveat of these sorting results is illustrated in figures 11 and 12, which show the traditional IV analyses results in the deciles of the IV GRF predicted effects for hospitalizations and spending respectively. As opposed to the first stage effect, the IV effects do not show a clear pattern as predicted IV effect increases. This suggests that predicting which encounters will benefit from ND care is more difficult than predicting which enocunters *will* shift to NDs.

7 Conclusion

Proponents of increased use of NDs argue that NDs do not produce worse quality care than MDs, and that NDs do not increase (and more likely decrease) patient medical spending. Instead, when I focus on the relevant sample of patients most likely to be affected by relaxing restrictions on NDs, I find an increase in preventable hospitalizations with noisy increases in spending. Specifically, I find that ND use

increases the share of encounters that were followed by a preventable hospitalization by roughly 0.5 percentage points. In other words, roughly 1 in 200 encounters is followed by a preventable hospitalization that would not have been had MDs handled the encounter. With a conservative estimate of the cost of an inpatient admission at roughly $10,000, this suggests that ND utilization essentially increases the cost of outpatient visits by roughly $50 each.

These conclusions are based on a sample of privately insured patients who are relatively well employed. Results are likely to be different on a set of patients with different outside options for care, such as Medicaid enrollees or rural patients who may forgo care entirely if they do not get access to NDs. Nevertheless, the set of patients studied in this paper are generally lower risk patients than the general population, and so increases in preventable hospitalizations in this group may be exacerbated in other more risky groups.

I find that relaxing SOP laws has potential to improve patient outcomes, dependent on the efficiency of sorting tasks between provider types. There is a large collection of encounters for which NDs provide improved outcomes over MDs, but those are not necessarily the types of encounters most likely to be shifted to NDs after the law changes. The sorting isn't as bad as it could be either: the types of encounters for which NDs provide the worst outcomes relative to MDs are also not the most likely to be shifted to NDs.

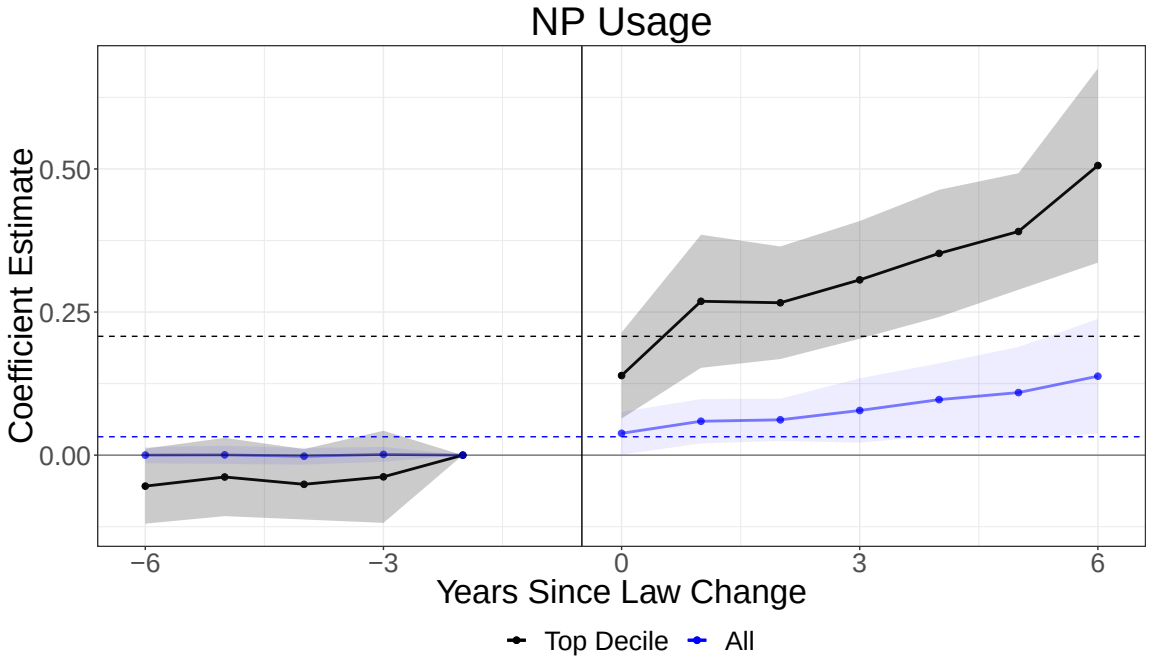
Another potential mechanism driving the effects is the changes in prescriotion drug behavior in the high first stage sample. Patients in this group decreased medication adherence rates and did not increase their prescription fill rates, while the overall sample did not decrease adherence rates but did increase their fill rates.

There are multiple interacting incentives that must be aligned in order to reach efficient sorting. First, patients or schedulers must be able to accurately assess their "riskiness" or the approximate value added of NDs so they can be sorted to provider type which will provide the best quality care. Furthermore, insurance companies must provide financial incentives to direct patients to the proper provider type. Skewed out-of-pocket costs could induce patients to rarely select the most cost effective provider type. Future work investigating the effects of these incentives

could provide valuable insight into the policy changes needed to optimize sorting of patients between provider types.

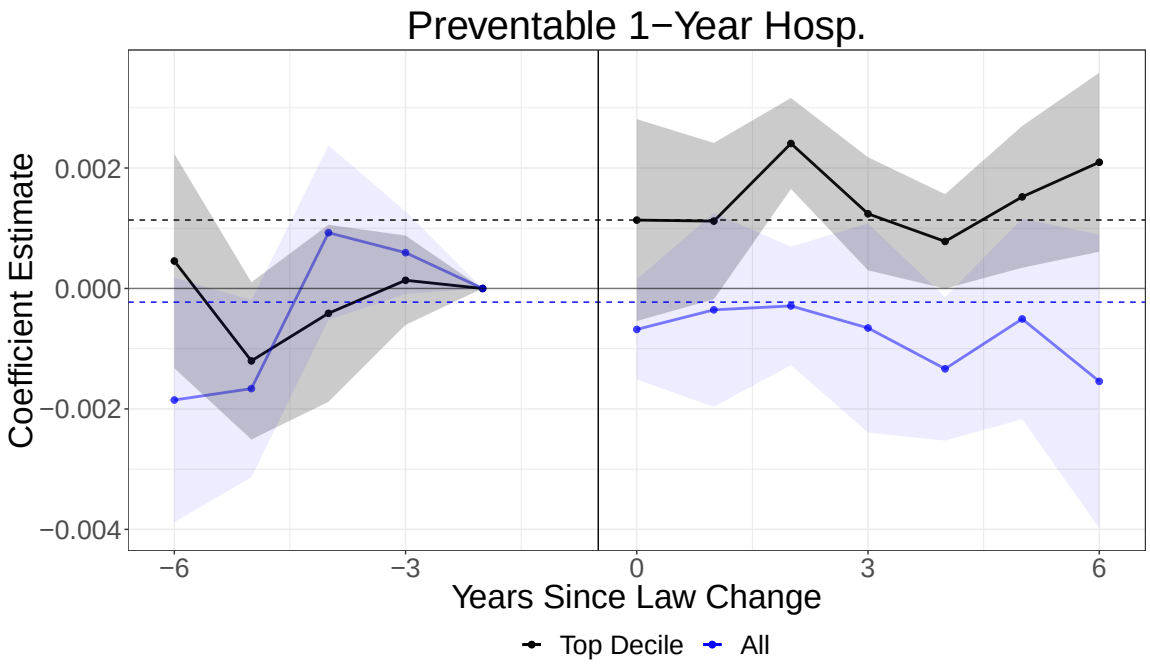
8 Tables and Figures

Figure 1: First Stage Event Study



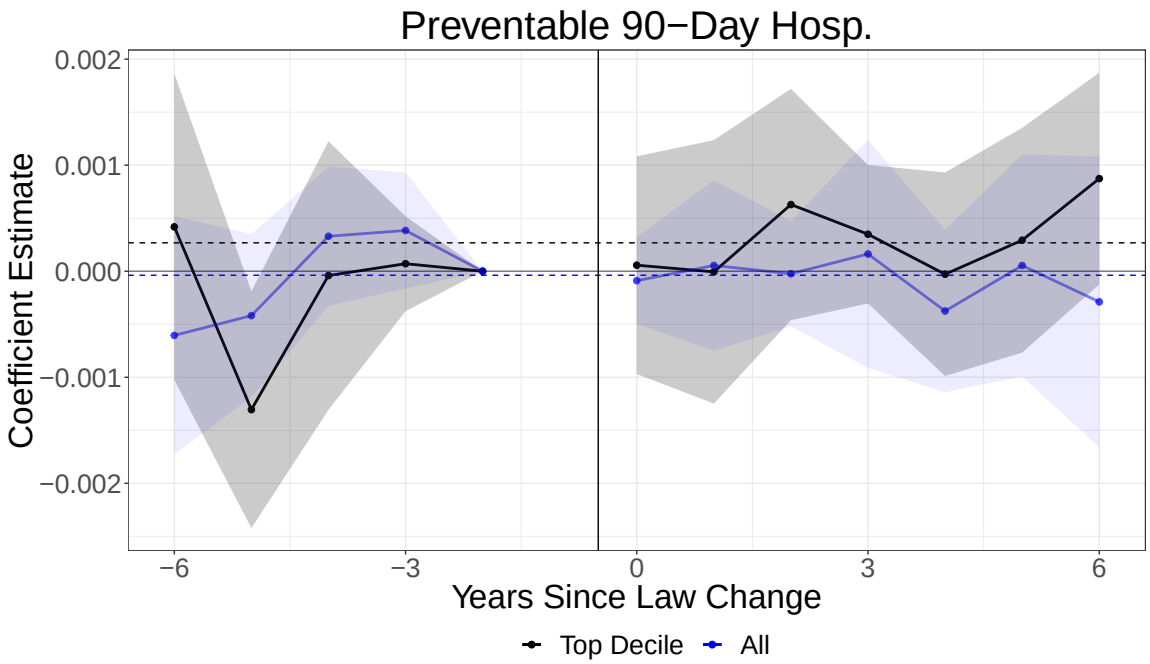
Note: This figure shows the coefficients on years since the SOP law change indicators in the first stage event study, where the outcome is an indicator for whether the encounter was handled by an ND. The regression also includes state and year fixed effects, as well as a linear time trend specific to states that changed their laws after the passing of the Affordable Care Act. The shaded area represents 95 percent confidence intervals based on standard errors clustered at the state level. The results from the full sample are shown in blue, while the results from the top decile of predicted first stage effects are shown in black.

Figure 2: Reduced Form Event Study (1-year Preventable Hospitalizations)



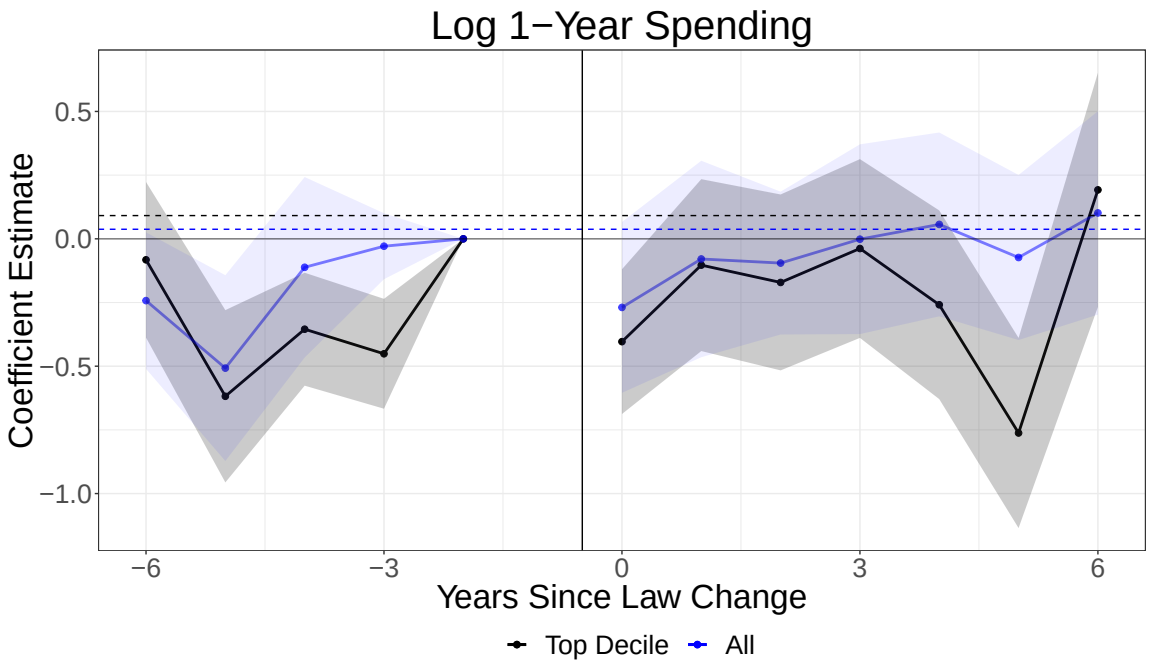
Note: This figure shows the coefficients on years since the SOP law change indicators in the reduced form event study where the outcome is an indicator for whether the encounter was followed by a hospitalization for a preventable reason within 1 year. The regression also includes state and year fixed effects, as well as a linear time trend specific to states that changed their laws after the passing of the Affordable Care Act. The shaded area represents 95 percent confidence intervals based on standard errors clustered at the state level. The results from the full sample are shown in blue, while the results from the top decile of predicted first stage effects are shown in black.

Figure 3: Reduced Form Event Study (90-day Preventable Hospitalizations)



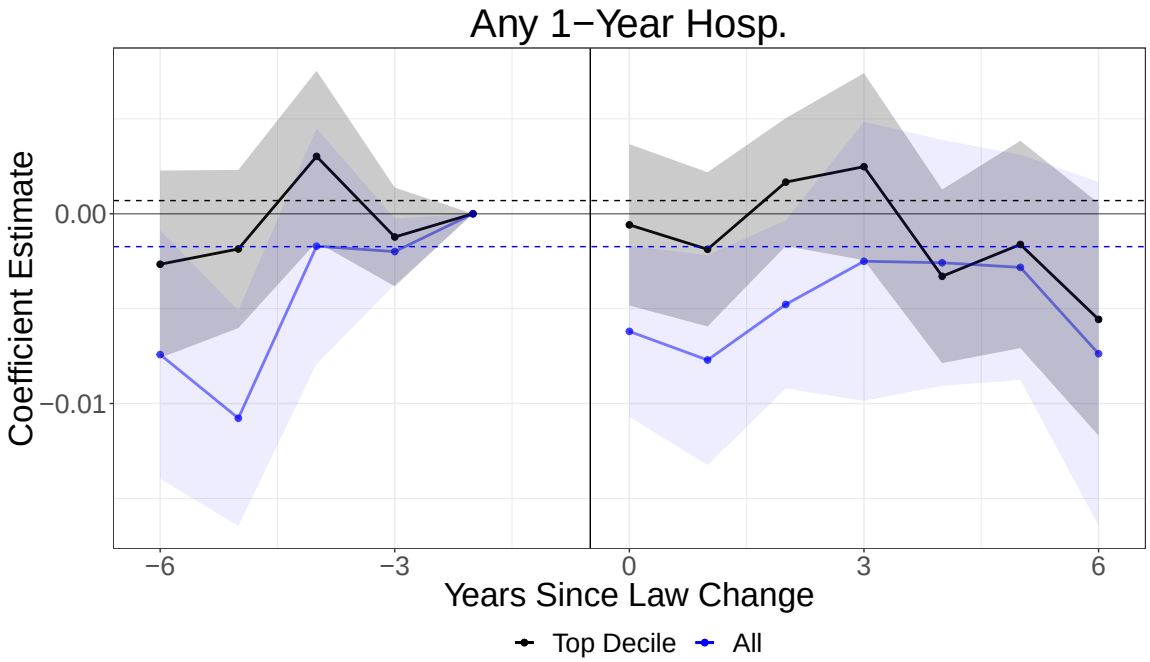
Note: This figure shows the coefficients on years since the SOP law change indicators in the reduced form event study where the outcome is an indicator for whether the encounter was followed by a hospitalization for a preventable reason within 90 days. The regression also includes state and year fixed effects, as well as a linear time trend specific to states that changed their laws after the passing of the Affordable Care Act. The shaded area represents 95 percent confidence intervals based on standard errors clustered at the state level. The results from the full sample are shown in blue, while the results from the top decile of predicted first stage effects are shown in black.

Figure 4: Reduced Form Event Study (Log 1-year Spending)



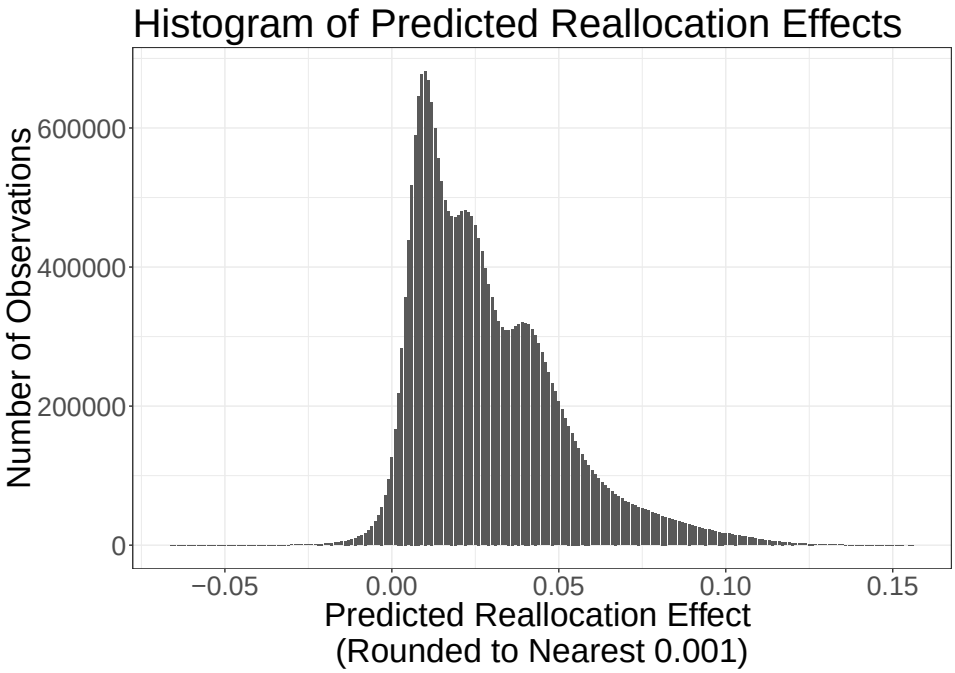
Note: This figure shows the coefficients on years since the SOP law change indicators in the reduced form event study where the outcome is the patient's total medical spending in the year following the encounter. The regression also includes state and year fixed effects, as well as a linear time trend specific to states that changed their laws after the passing of the Affordable Care Act. The shaded area represents 95 percent confidence intervals based on standard errors clustered at the state level. The results from the full sample are shown in blue, while the results from the top decile of predicted first stage effects are shown in black.

Figure 5: Reduced Form Event Study (1-year Any Hospitalizations)



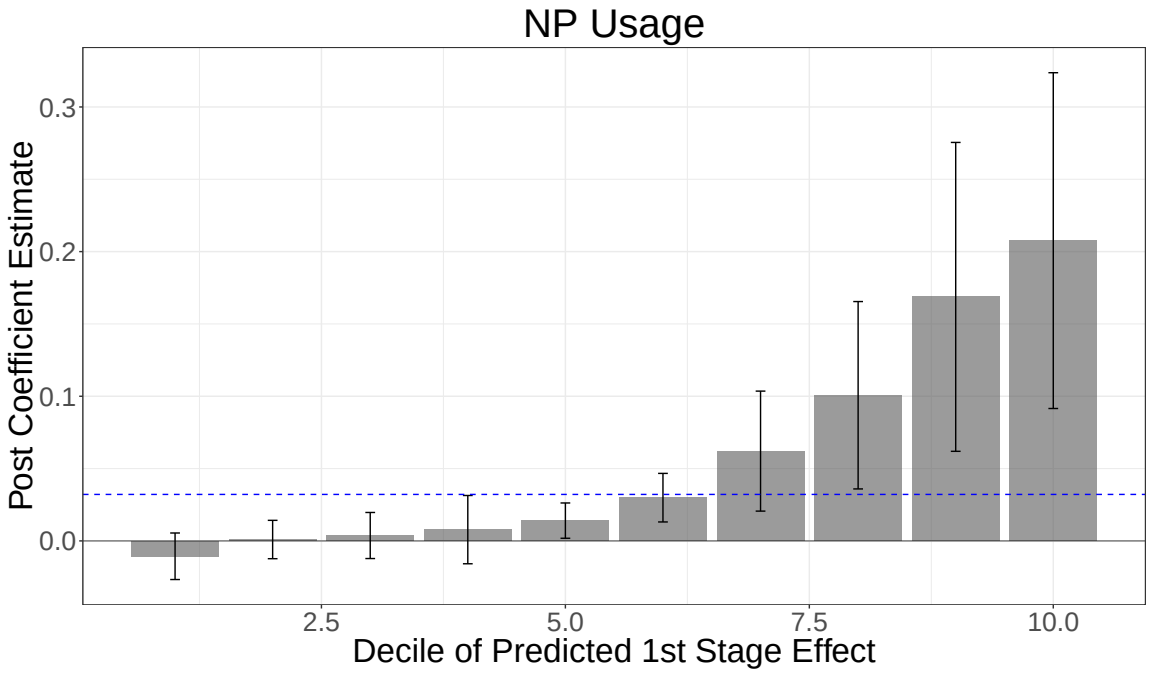
Note: This figure shows the coefficients on years since the SOP law change indicators in the reduced form event study where the outcome is an indicator for whether the encounter was followed by a hospitalization for a any reason within 1 year. The regression also includes state and year fixed effects, as well as a linear time trend specific to states that changed their laws after the passing of the Affordable Care Act. The shaded area represents 95 percent confidence intervals based on standard errors clustered at the state level. The results from the full sample are shown in blue, while the results from the top decile of predicted first stage effects are shown in black.

Figure 6: Predicted Allocation Effect Distribution



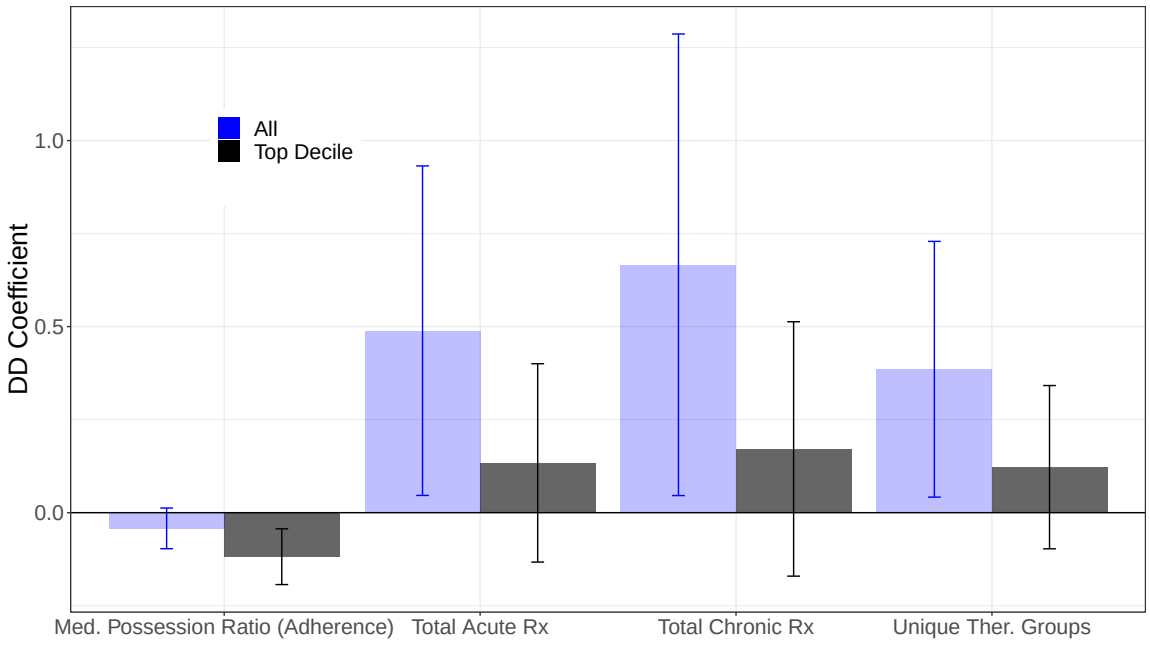
Note: This figure shows the distribution of predicted first stage allocation effects.

Figure 7: ND First Stage Effect by Predicted First Stage Decile



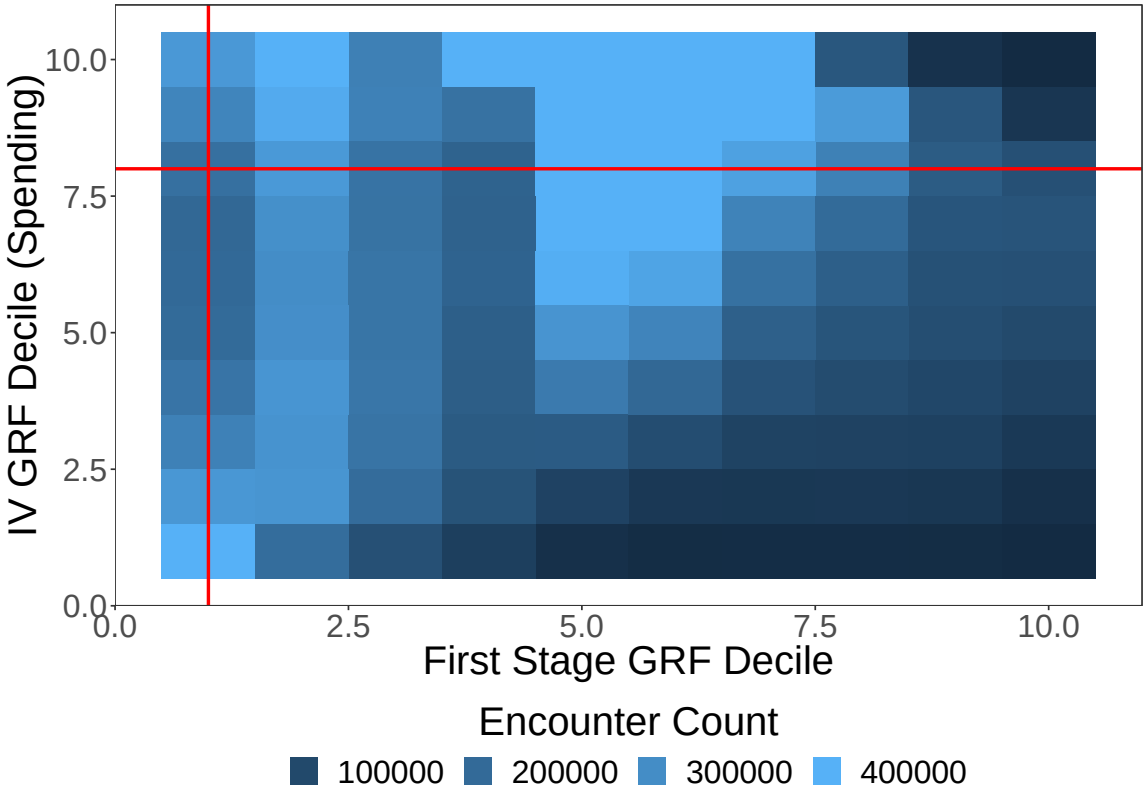
Note: This figure shows the DD coefficient on the post law change indicator in specification with the ND use indicator as the outcome. The bars show the effect in each decile of predicted first stage effects. Error bars show 95% confidence intervals based on standard errors clustered at the state level. The blue dashed line shows the effect from the traditional analysis in the overall sample.

Figure 8: RX Effect Breakdown



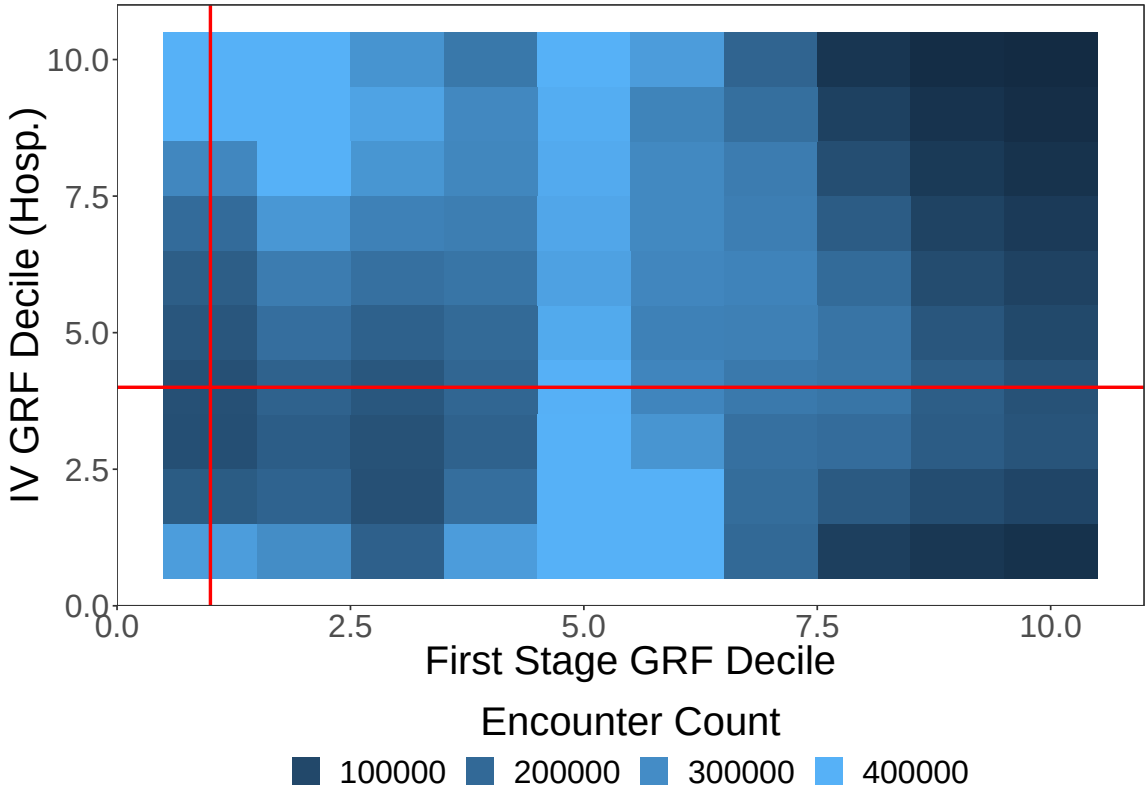
Note: This figure shows the DD coefficient on the post law change indicator in specification with the listed variable as the outcome. The blue bars represent the effect in the overall sample, while the black bars show the effect in the top decile of first stage effects. Error bars show 95% confidence intervals based on standard errors clustered at the state level.

Figure 9: Compare First Stage GRF to IV GRF (1-Year Log Spending)



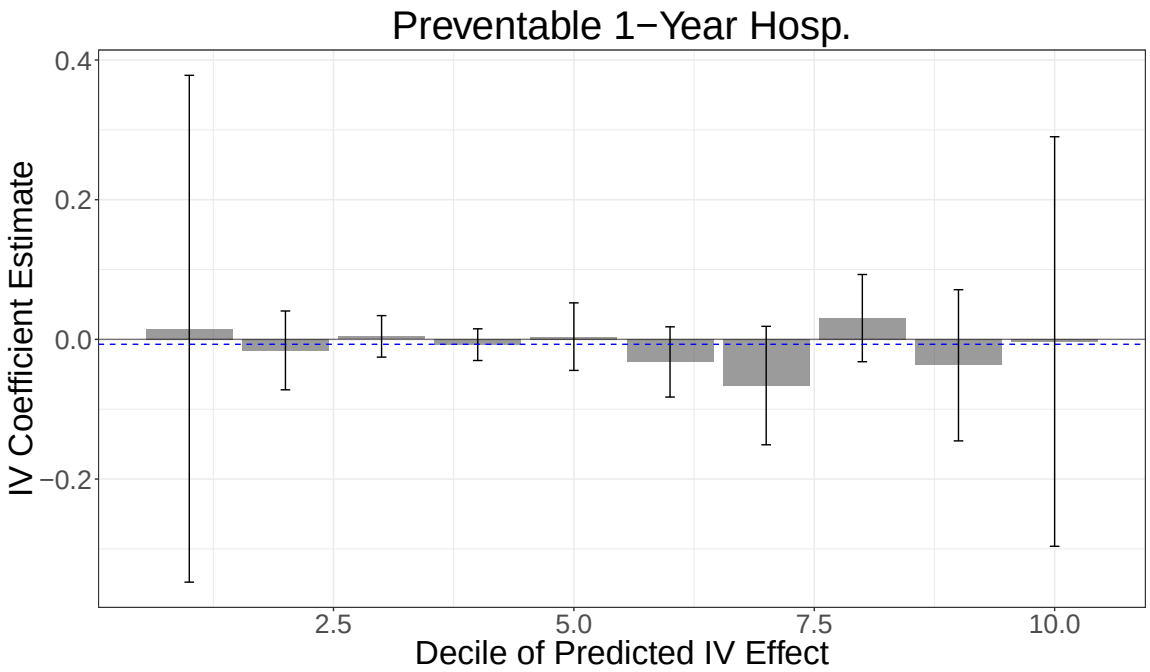
Note: This figure shows a heatmap of the number of encounters in each first stage GRF decile - IV GRF decile bin. Light colors indicate more unique patients. The red lines indicate the decile at or below which predicted effects are negative (so the first stage GRF predicted effects are positive in all but the first decile, and the IV GRF predicted effects are positive in all deciles above the eighth decile).

Figure 10: Compare First Stage GRF to IV GRF (1-Year Hospitalizations)



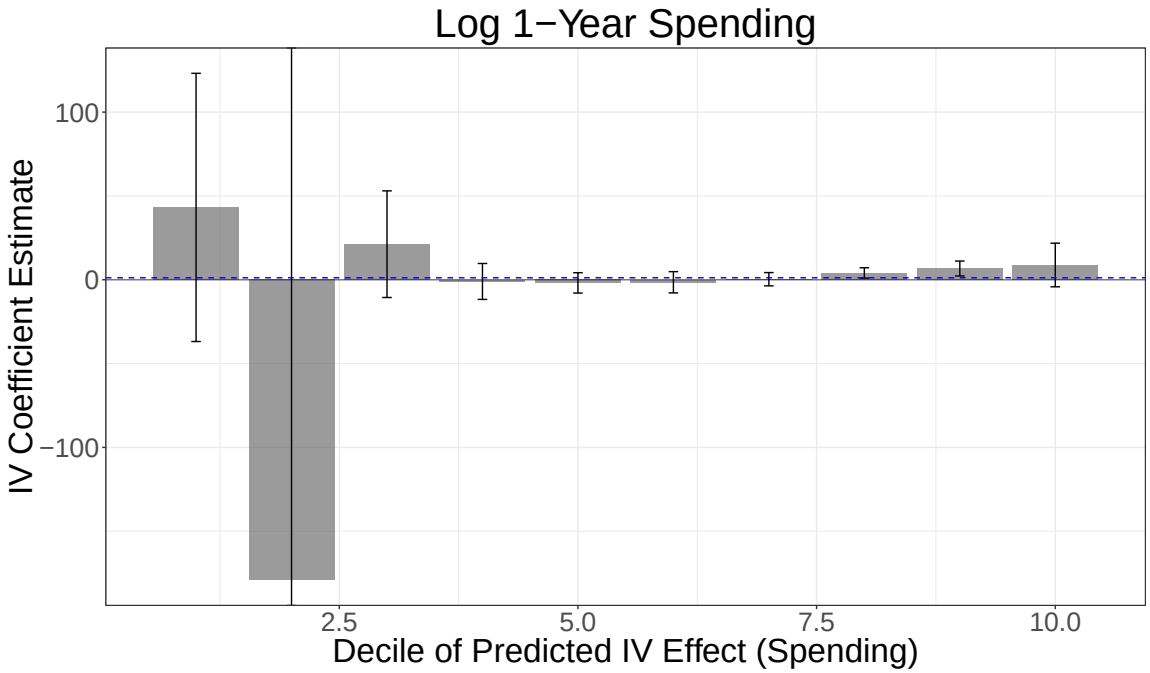
Note: This figure shows a heatmap of the number of encounters in each first stage GRF decile - IV GRF decile bin. Light colors indicate more unique patients. The red lines indicate the decile at or below which predicted effects are negative (so the first stage GRF predicted effects are positive in all but the first decile, and the IV GRF predicted effects are positive in all deciles above the fourth decile).

Figure 11: IV Effects (1-year Preventable Hospitalizations)



Note: This figure shows the coefficients on years since the SOP law change indicators in the reduced form event study where the outcome is an indicator for whether the encounter was followed by a hospitalization for a any reason within 1 year. The regression also includes state and year fixed effects, as well as a linear time trend specific to states that changed their laws after the passing of the Affordable Care Act. The shaded area represents 95 percent confidence intervals based on standard errors clustered at the state level. The results from the full sample are shown in blue, while the results from the top decile of predicted first stage effects are shown in black.

Figure 12: IV Effects (1-year Log Spending)



Note: This figure shows the coefficients on years since the SOP law change indicators in the reduced form event study where the outcome is an indicator for whether the encounter was followed by a hospitalization for a any reason within 1 year. The regression also includes state and year fixed effects, as well as a linear time trend specific to states that changed their laws after the passing of the Affordable Care Act. The shaded area represents 95 percent confidence intervals based on standard errors clustered at the state level. The results from the full sample are shown in blue, while the results from the top decile of predicted first stage effects are shown in black.

Table 1: State Law Changes

Change DURING Data	Year of Change	Change BEFORE Data	Change AFTER Data	
CO	2010	AK	AL	MO
CT	2014	AZ	AR	MS
HI	2011	DC	CA	NC
ID	2004	IA	DE	NY
MD*	2010	ME	FL	OH
MN	2015	MT	GA	OK
ND	2011	NH	IL	PA
NE	2015	NJ	IN	SC
NV	2013	NM	KS	SD
RI	2008	OR	KY	TN
VT	2011	UT	LA	TX
WY	2005	WA	MA	VA
		WV	MI	WI

Notes: Year of change years represent the years the states changed from supervized or collaboratory agreements between NDs and MDs to independence of practice for NDs. * I exclude Maryland from the estimation sample due to ambiguity in the law change date.

Table 2: Summary Statistics, Overall Sample

Variable	Mean/(SD)	Variable	Mean/(SD)
Age	35.516	NP Usage	0.034
	(16.404)		(0.182)
Male	0.406	Any Hosp Next Year	0.058
	(0.491)		(0.233)
Rural	0.13	Preventable Hosp Next Year	0.006
	(0.337)		(0.08)
Overall Copay	4.94	Preventable Hosp Next 90 Days	0.003
	(11.836)		(0.05)
Overall Payment	96.099	Total Spending Next Year	6,645.522
	(514.548)		(28967.083)
Total Encounters	24,334,637	Log Total Spending Next Year	6.795
Total Patients	1,388,886		(2.684)

Note: This table shows the means and standard deviations for the variables listed, as well as the total counts of encounters and patients in the overall sample.

Table 3: Summary Statistics, High First Stage Sample

Variable	Mean/(SD)	Variable	Mean/(SD)
Age	22.749	NP Usage	0.092
	(15.407)		(0.289)
Male	0.32	Any Hosp Next Year	0.01
	(0.466)		(0.1)
Rural	0.181	Preventable Hosp Next Year	0.002
	(0.385)		(0.039)
Overall Copay	4.41	Preventable Hosp Next 90 Days	0.001
	(9.899)		(0.028)
Overall Payment	65.126	Total Spending Next Year	921.7484
	(249.433)		(6176.96)
Total Encounters	672,408	Log Total Spending Next Year	3.743
Total Patients	138,889		(3.07)

Note: This table shows the means and standard deviations for the variables listed, as well as the total counts of encounters and patients in the large predicted first stage subsample.

Table 4: Regression Coefficients, Overall Sample

	Baseline Mean	OLS	IV
NP Usage	0.024	0.032	
	(0.153)	(0.012)	
Log 1-Year Spending	6,924.23	0.038	1.169
	(26,331.51)	(0.1004)	(3.143)
Preventable 1-Year Hosp.	0.007	-0.0002	-0.007
	(0.084)	(0.0005)	(0.015)
Preventable 90-Day Hosp.	0.003	-0.00004	-0.001
	(0.052)	(0.0003)	(0.008)
Any 1-Year Hosp.	0.063	-0.002	-0.054
	(0.243)	(0.002)	(0.05)
N		24,334,637	24,334,637
First Stage F			7.341

Note: This table shows the regression output for the reduced form and IV specifications in the overall sample. Each row represents a separate regression with the outcome variable listed. The first column (labeled "Baseline Mean") shows the baseline (pre-law change) means and standard deviations. The second column shows the reduced form effects of the law change on the outcome variable. The final column shows the IV coefficient on the ND use indicator, instrumented by the post law change indicator. All specifications include state and year fixed effects, as well as a linear time trend specific to states that changed their laws after the passing of the Affordable Care Act (2010). Standard errors, clustered at the state level, are shown in parentheses. For the IV specifications, The first stage F-statistic is shown at the bottom of the table.

Table 5: Regression Coefficients, High First Stage Sample

	Baseline Mean	OLS	IV
NP Usage	0.065	0.208	
	(0.246)	(0.059)	
Log 1-Year Spending	859.49	0.091	0.439
	(5,037.18)	(0.2072)	(0.954)
Preventable 1-Year Hosp.	0.002	0.0011	0.005
	(0.041)	(0.0003)	(0.002)
Preventable 90-Day Hosp.	0.001	0.00027	0.00129
	(0.03)	(0.0003)	(0.001)
Any 1-Year Hosp.	0.011	0.001	0.003
	(0.104)	(0.001)	(0.005)
N		672,408	672,408
First Stage F			12.29

Note: This table shows the regression output for the reduced form and IV specifications in the high first stage subsample. Each row represents a separate regression with the outcome variable listed. The first column (labeled "Baseline Mean") shows the baseline (pre-law change) means and standard deviations. The second column shows the reduced form effects of the law change on the outcome variable. The final column shows the IV coefficient on the ND use indicator, instrumented by the post law change indicator. All specifications include state and year fixed effects, as well as a linear time trend specific to states that changed their laws after the passing of the Affordable Care Act (2010). Standard errors, clustered at the state level, are shown in parentheses. For the IV specifications, The first stage F-statistic is shown at the bottom of the table.

A Data Appendix

This section explains details about the creation of the estimation data sample.

To avoid changes in outcome measures driven by patients "aging out" of the sample into retirement and/or medicare, I limit to patients who are younger than 60 years old throughout the entire duration of the study.

Since some encounters with multiple providers may occur on the same patient "visit" where a patient is billed both for the ND encounter and the MD encounter, I limit to cases where patients were either seen by an ND and not seen by an MD on the same day, or seen by an MD and not seen by an ND on the same day.

A common occurrence in the claims data is that claims will be "adjusted" after initial submission by the insurance agency or the medical providers. For example, suppose an initial claim is filed with payment amount Z. After negotiations between the insurer and the medical provider, the payment amount is negotiated to a different amount X = Z + Y, where Y is the "adjustment" between the two submitted claims and may be negative. To reflect this alteration, another claim is filed with the identical medical information, but with a new payment amount Y. This process may continue for several iterations over the same encounter. I handle these adjustments by collapsing the claims so that each encounter is only represented once, with the payment amount equal to the sum of the payment variables from each of the claims submitted regarding the encounter. Thus, each observation in the data is an encounter with payment equal to the final amount paid by or on behalf of the patient by his or her insurance agency.

The estimation sample is a subset of all patients in the full data, created in the following way. First, I limited to patients in states which changed their SOP laws between 2004 and 2015, who had at least one medical encounter with a general practice physician or a nurse practitioner. Then I took the full history of all claims for this group of patients. I use the full history for 25 percent of these patients for the estimation, and a different 25 percent for the training of the machine learning algorithm. The remaining 50 percent of the patients are reserved for future robustness analyses. I include all outpatient claims, all inpatient claims, and all prescription

drug claims which constitute my three master data sets.

I then collapse each of the data sets to correct for payment adjustments as described above, and create a new data set that contains just the patient identification number and the dates of encounters with general practice physicians or nurse practitioners. I use this as the skeleton for the final estimation sample, merging on outcome variables created using all three master data sets (total medical spending in the year following the MD/ND visit by or on behalf of the patient, and whether or not the MD/ND visit was followed by a preventable hospitalization in the next year). I also merge on control variables created from the three master data sets, such as lagged information about spending, the total unique number of diagnostic categories, the total number of hospitalizations, ER visits, and prescription drug claims; as well as information about the MD/ND visit including patient age and gender, major diagnostic category, facility type, and type of insurance plan. Thus, the result is a data set where each observation is an encounter with an MD or an ND, and each column is a characteristic of that encounter, previous encounters by the patient, or future medical information. I use the same set of patients to estimate the overall traditional econometric results as well as the final HTE estimates.

B GRF Appendix

The generalized random forest (GRF) is a forest-based weighting algorithm that partitions data by using recursive axis-aligned splits to maximize the heterogeneity of the treatment effect estimates between partitions. This process is repeated across subsamples of the data to create a set of partition rules, which are used to create a vector of weights that establishes the similarity of any two observations. This weighting vector is used to solve a weighted local maximum likelihood optimization problem which provides an estimate for the treatment effect for each observation.

The specific process begins by taking a sample I from the full training data set, where I is s percent of the training data. I cluster at the patient level, so each sample I contains all information for each of the patients in that sample. I is then further split into two equal sized subsets J_1 and J_2, again clustered at the patient

level. Subset J_1 is used to grow a regression tree T (see process explanation below), which returns a partitioning rule that fully partitions data J_1 into terminal nodes or "leaves". Next, the algorithm applies the partitioning rule from T to the other subsample J_2, so that J_2 is fully partitioned into leaves. I repeat this process until I reach the total number of trees, B, resulting in B partitioning rules on B (potentially overlapping) subsets of the training data, each of size $s/2$. The partitioning rules are based on the independent variables of each observation.

Then I turn to the testing data set. For each observation x in the testing data set, I "insert" x into each of the B partitioning rules so that x is in B leaves each of size N_b, where b indexes the specific partitioning rule. For each rule $b = 1, \ldots, B$ I assign a weight equal to $1/N_b$ to all points in the set of points in the corresponding J_2 subsample that are in the same leaf as x. Thus, each observation in the training data has a vector of weights of length B corresponding to the single observation x in the testing data. I then take the mean of all B weights for each observation in the training data, and use the means as the final weight vector, which I then use to solve a local maximum likelihood problem to estimate a treatment effect estimate specific to the observation x. I then repeat this process for each observation in the testing data set, so that each observation in the testing data has a treatment effect estimate.

To grow each tree, the algorithm first proposes a split based on the independent variables (such as $Age > 40$). For each proposed split, I solve a local estimating equation in both of the resulting child nodes. The estimating equation is a local maximum likelihood problem which gives a treatment effect estimate in each child node. The proposed split with the largest difference between in treatment effect estimates between the child nodes is selected as the actual split. The process is then repeated on each of the child nodes, until either of the resulting nodes has no treatment group or control group observations or any proposed split does not increase the heterogeneity in treatment effect estimates between the two child nodes.

To understand which variables drive the prediction estimates, I regressed the first stage predictions on the predictor variables, without any non-linear terms or interaction terms. This approach misses an important feature of the GRF, namely

its ability to capture non-linearities and interactions in the data, but the linear approach can give an approaximation for each variable's explanatory power. I show the 20 variables with the smallest p-values in the regression in the right hand column of table 1 (with the lowest p-values at the top and the largest at the bottom). The left column of table 1 shows the least important predictor variables (as determined by the p-values). Patient employment industry, age, and gender are important predictors of switching to NDs. Facility type and some (but not all) specific information about a patient's medical history are the least important predictors.

Table 1: Variable Importance

Least Significant	Most Significant
mdc 3 last year	outpatient unique mdc 8 count last year
thergrp 31 count last year	outpatient unique mdc 19 last year
any hosp last 3 years	deduct
thergrp 7 last year	mdc8
preventable hosp last 3 years	industry2
outpatient unique spec 430 count last year	tot outpatient spending last year
ER mdc 0 count last year	outpatient unique mdc 9 last year
ER mdc 3 count last year	monthly mean outpatient spending last year
mdc 14 count last year	mdc3
thergrp 30 count last year	industry3
mdc 22 last year	outpatient unique mdc 8 last year
ER mdc 10 count last year	industry5
outpatient unique spec 575 count last year	outpatient unique mdc 3 last year
(Intercept)	tot outpatient spending spike last year
facility type18	male
facility type95	in network
facility type98	outpatient unique mdc 23 last year
facility type42	industry6
facility type49	AGE
facility type35	industryNA

Note: This table shows the 20 least "important" variables (left column) and the 20 most "important" variables (right column). Importance here is determined by the p-values in a linear regression of the first stage GRF prediction on all predictor variables.