NBER WORKING PAPER SERIES

SCREENING THROUGH SOFT SPENDING LIMITS: EVIDENCE FROM THE MEDICARE THERAPY CAP

Ashvin Gandhi
Maggie Shi

Working Paper 33722
http://www.nber.org/papers/w33722

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2025, Revised January 2026

We thank Zhijian Li, Katherine Papen, and Fuman Xie for excellent research assistance, as well as Jason Falvey, DPT; Brian McGarry, PhD, PT; Patti Marquardt, PT; and Xiao Zheng, MD for sharing their clinical expertise. We thank Zarek Brot, Samantha Burn, Josh Gottlieb, Tim Layton, Riley League, Jetson Leder-Luis, Ryan McDevitt, Dan Sacks, Jacob Wallace, and seminar participants at ASHEcon, the Junior Health Economics Workshop, the Toulouse School of Economics, IU Bloomington, Chicago Health Economics Workshop, the Berkeley Health Economics Workshop, UIUC, the Hoover Institute, UC Santa Barbara, Stanford, BFI Health Conference, NBER SI Economics of Health meeting, Emory, the Midwest Health Economics Conference, Monash University, the American Health Econometrics Workshop, the Southeastern Health Economics Study Group, the Penn Leonard Davis Institute, UChicago Harris, Bowdoin College, and the Congressional Budget Office for helpful comments. The authors gratefully acknowledge support from the National Institute on Aging (#T32-AG000186) and Arnold Ventures. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Screening Through Soft Spending Limits: Evidence from the Medicare Therapy Cap
Ashvin Gandhi and Maggie Shi
NBER Working Paper No. 33722
April 2025, Revised January 2026
JEL No. H30, H51, I1, I13, I18

ABSTRACT

Governments and firms often use soft spending limits to curb overspending while allowing exceptions. We study a Medicare policy that capped per-patient physical therapy spending with exceptions for documented medical need. By screening out low-value care, the cap reduced spending by 8 percent without harming patient health. This screening was driven by Medicare's exception decisions, whereas patients and providers scaled back care indiscriminately. However, the cap also generated horizontal inequities: conditional on need, lower-income and minority patients were more likely to be screened out because they were treated by smaller providers who were slower to adapt to new documentation requirements.

Ashvin Gandhi
University of California, Los Angeles
Anderson School of Management
and NBER
ashvin.gandhi@anderson.ucla.edu

Maggie Shi
University of Chicago
Harris School of Public Policy
and NBER
m.shi@uchicago.edu

1 Introduction

Governments and firms frequently delegate spending decisions to agents, but these agents may choose to spend in ways the principal deems wasteful. One way to mitigate this is to require agents to seek approval for their spending by submitting documentation to justify it. When these requirements apply above a stated threshold, they form a "soft" spending limit in that it can be bypassed on a case-by-case basis. For example, government procurement policies often require documented justification to exceed statutory cost caps (CFR, 2020). More broadly, requiring documented justification for approval to receive *any* funds—effectively placing a soft limit at zero—is a common feature of many social programs, including disability benefits (SSA, 2008), unemployment insurance (DOL, 2025), and disaster assistance (FEMA, 2021).

These soft screening mechanisms aim to curb wasteful spending while still allowing flexibility for exceptional needs. They can improve efficiency by screening in two ways: the documentation informs the principal's decision to *deny* or approve requests, and the ordeal can lead agents to "self-screen" by *deterring* low-value requests (Nichols and Zeckhauser, 1982). However, they may also inadvertently generate horizontal inequity by favoring agents with greater administrative capacity: those better at producing documentation may face fewer denials and also be less deterred. Despite the widespread use of soft screening mechanisms, there has been little empirical study unpacking these efficiency and equity considerations.

This paper studies these considerations in the context of a soft limit on healthcare spending. While initial care decisions are delegated to patients and providers, an insurer usually has the final say in approving reimbursement for care. One way insurers mitigate overspending is by requiring providers to submit documentation of medical need for all care above the spending limit. This practice is especially common for treatments particularly susceptible to moral hazard, like physical therapy, psychotherapy, chiropractic, and dental services.[1]

We study a soft spending limit imposed on physical therapy (PT) within the Medicare program, which was established in response to concerns about medically unnecessary overuse of services like manual massage and supervised strength training. The "therapy cap" was an annual, per-patient spending limit of $1,740, or about 11 weeks of care. Providers could request exceptions for individual patients with documented need to exceed the limit. Medicare then approved or denied each request based on their assessment of medical necessity.

This setting is well-suited for understanding how soft screening mechanisms work more generally. The policy we study imposes a sharp threshold on a recurring service, which

[1] In the extreme, insurers can place restrictions on the *first* unit of care, a practice known as prior authorization (Dillender, 2018; Brot-Goldberg et al., 2023; Burn and Ristovska, 2025).

allows us to distinguish between deterrence—patients who independently stop just short of the cap—and denials—patients who attempt to exceed it, but are stopped by Medicare. Furthermore, the rich data in this context includes measures of patient health, which allow us to evaluate the targeting of cap-induced savings. These data also include measures of documentation, which allow us to assess the role of administrative capacity in mediating these effects.

We first quantify the overall savings from the therapy cap with a difference-in-bunching estimator that compares the distribution of PT spending before and after the 2006 introduction of the cap. Overall, the therapy cap reduced spending by eight percent relative to pre-reform levels. We find that 58 percent of savings are due to Medicare denials, while the remaining 42 percent stems from deterrence.

We then employ two approaches to evaluate the screening properties of these spending reductions. The first looks directly for negative patient health effects that would be indicative of poorly targeted cutbacks. We leverage the fact that the cap resets each calendar year and is therefore more binding for patients who start PT earlier in the year. Using this variation, we find that cap-induced savings were well-targeted: they neither resulted in substitution to alternatives like opioids, pain procedures, and orthopedic surgeries, nor in worsening health as proxied by hospitalizations, emergency department visits, and nursing home stays.

Our second approach tests whether the patients screened out by the cap have lower observable clinical need. Importantly, here we can distinguish between screening through denials and screening through deterrence. We find that this distinction matters: while savings derive from both deterrence and denials, screening on need is driven entirely by Medicare's denial decisions. That denials screen on need is consistent with Medicare's stated objective of reducing unnecessary care. The lack of similar screening in deterrence, however, runs counter to the standard prediction of precise "self-screening" along the same dimension (Zeckhauser, 2021). This suggests that patients and providers respond to the ordeal associated with the cap fairly bluntly, reducing care across the board rather than in a targeted way.

We then consider whether the cap inadvertently screened along dimensions other than medical need: specifically, on providers' administrative ability to comply with the documentation requirements. Using provider size as a proxy for administrative ability, we find substantial disparities in denial rates for otherwise-similar patients. Conditional on medical need, patients of small providers are 55 percent (12.7 percentage points) more likely to be denied by the cap than those seeing large providers.

Notably, the differences in denial rates across providers lead to substantial horizontal inequities across key patient demographic groups: low-income and minority patients face higher denial rates because they are more likely to see small providers. Conditional on need,

non-white patients, Medicaid enrollees, and Part D Low Income Subsidy recipients are 20-38 percent more likely to be denied by the cap than their respective counterparts. These disparities are driven by patient sorting across providers, as they largely disappear with the inclusion of provider fixed effects.

Finally, we investigate the mechanisms that drive large providers' advantage with cap denials. We first show that getting approved to go past the cap is closely tied to whether the provider has supporting documentation available, and large providers are much more likely to provide documentation. However, this advantage in compliance is not fixed and appears to come from learning-by-doing: both large and small providers learn to provide appropriate documentation as they gain experience with the cap, but large providers mechanically accrue experience more quickly.

Our findings point to two distinct lessons about policies aimed at controlling costs. First, denials by the regulator are central to the efficacy of these screening mechanisms. While ordeals do deter agents from spending, they curtailed both high- and low-value spending to similar extents. Second, by implicitly screening on agents' administrative capacity, these policies can introduce horizontal inequity for expenditures with similar underlying value. The resulting inequities are particularly concerning when administrative capacity is correlated with traits that should be orthogonal to the value of the spending in question, like race.

This paper contributes to the literature on the effectiveness and targeting of screening mechanisms (Nichols and Zeckhauser, 1982; Kleven and Kopczuk, 2011; Alatas et al., 2016; Deshpande and Li, 2019; Finkelstein and Notowidigdo, 2019; Lieber and Lockwood, 2019; Ida et al., 2022; Rafkin et al., 2023; Shepard and Wagner, 2024) by being the first to separately evaluate the deterrence and denial channels. Making this distinction matters: while both channels contribute to savings, they have very different screening properties. Our specific application also contributes to a growing literature studying the efficacy of these tools in the context of curbing wasteful healthcare spending: related examples include prior authorization (Dillender, 2018; Brot-Goldberg et al., 2023; Eliason et al., 2024; Burn and Ristovska, 2025), audits or denials (Macambira et al., 2022; League, 2023; Shi, 2024), and anti-fraud enforcement (Howard and McCarthy, 2021; O'Malley et al., 2023; Leder-Luis, 2023).

Finally, our study speaks to the literature on disparities in takeup of public programs. While much of this literature has focused on beneficiary-facing barriers like information gaps and application costs (Currie, 2006), we highlight a largely unexamined factor: the intermediaries who handle paperwork on behalf of beneficiaries. We find that heterogeneity in their administrative ability leads to substantial variation in access, resulting in meaningful horizontal inequity across beneficiaries with identical observable need.[2]

[2]Providers submit claims and handle documentation on behalf of their patients. In this way, their role is

2 Policy Context and Data

2.1 Outpatient Physical Therapy Care in Medicare

In 2017, Medicare Part B spent $4.9 billion on outpatient physical therapy services for 2.6 million beneficiaries, or about 7 percent of all traditional Medicare beneficiaries. Medicare beneficiaries can receive three main types of therapy in the outpatient setting: physical therapy (PT), occupational therapy (OT), and speech-language pathology (SLP).

A patient is typically referred to PT by a physician, and the physical therapist must develop a plan of care for the patient, which the physician certifies. The plan of care includes a diagnosis, long-term treatment goals, and the type, quantity, duration, and frequency of therapy services. Outpatient Medicare PT services are primarily provided in private practices (34%), nursing facilities (38%), and hospitals (15%) (APTA, 2020). Regardless of setting, outpatient PT is subject to the standard Medicare Part B cost-sharing rules: for approved services, Medicare pays 80 percent of allowed charges and the patient is responsible for the remaining 20 percent.

Patients might like to continue treatments such as directed exercise and massage even after their clinical benefits do not justify the costs for Medicare (OIG, 2016, 2017). As such, policymakers have long been concerned about medically unnecessary overuse of therapy services in the Medicare program. In 1993, the predecessor to the Centers for Medicare and Medicaid Services (CMS) launched a task force to address widespread reports of overbilling for therapy (US GAO, 1996). In the years that followed, CMS tried many policies to curb therapy spending, including the spending limits we study. An Office of the Inspector General (OIG) audit found that a third of Medicare outpatient therapy claims were still for medically unnecessary services, and over half of claims reviewed were not compliant with medical necessity, coding, or documentation requirements (OIG, 2018). According to the OIG, the most common reason for unnecessary PT services was patients receiving excessive amounts of therapy—that is, although a patient may have initially needed therapy, the amount, frequency, or duration they received went beyond "standards of practice."[3] Thus, efforts to curb excessive PT spending have focused on limiting the amount of per-patient spending in the form of annual "therapy caps."

analogous to that of tax preparers (Kopczuk and Pop-Eleches, 2007; Zwick, 2021), disability lawyers (Hoynes et al., 2022), social work case managers (Evans et al., 2024), and mortgage brokers (Woodward and Hall, 2012).

[3]Medicare requires the following for a therapy service to be considered "reasonable and necessary": (1) that the services are an effective treatment for a patient's specific condition, (2) that the service must be performed (or supervised) by a therapist, (3) that the service is expected to improve a patient's condition or is necessary to maintain their condition, and (4) that the amount, frequency, and duration of therapy follow standards of practice (CMS, 2020).

2.2 Medicare Therapy Cap

Medicare's policies on physical therapy spending can be divided into three regimes: a one-year "hard cap" in 1999, a six-year period with effectively no cap, and a period with a "soft cap" beginning in 2006 that continues to today. We focus primarily on the 2006 soft cap, but discuss the 1999 hard cap in Appendix Section C. The soft cap was first implemented as two $1740 caps in January 2006 — one placed on PT and SLP, and another placed on OT. We limit our analysis to the PT/SLP cap given that PT accounts for the majority of outpatient therapy spending. Medicare introduced a process through which providers could request exceptions for medically necessary services above the cap, thus making it "soft." When billing for services that would push a patient above the cap, they were supposed to indicate that they had documentation justifying medical necessity by using a billing code called the "KX" modifier code (CMS, 2006a). Providers did not have to attach the documentation to their claim; instead, using the modifier indicated that they "attested" that the documentation indicated that spending above the cap was "reasonable and necessary," and was available should Medicare request it. All attempts to bill over the cap were supposed to include this modifier code, though enforcement of this rule was inconsistent—16 percent of claims in which the modifier documentation should have been used but was not were still paid out in full in 2006.

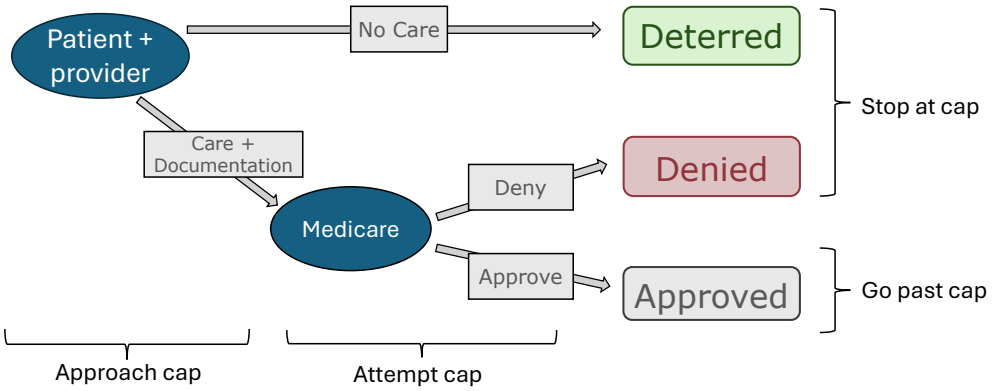
In using the documentation modifier code, the provider attested that the following documentation was available for review by Medicare: evaluation and plan of care, certification and re-certifications with evidence of physician (or non-physician practitioner) approval, progress reports, treatment notes, as well as potentially separate justifications for services that are "more extensive than is typical for the condition treated" (CMS, 2024a). The documentation is expected to justify that the patient requires the skill of a therapist, that the services are the appropriate type, frequency, intensity, and duration for the particular needs of the patient, as well as provide relevant information about the patient's functional abilities (CMS, 2006a).

The cap was enforced by Medicare through claim denials, as shown by the increase in denial rates at the cap that appears in 2006 in Figure B1. In order for the therapist to charge the patient for care in the event of a Medicare denial, she must issue an "advance beneficiary notice" (ABN) *prior* to the delivery of the service. The ABN informs the patient why Medicare may not pay for a specific claim and allows them to choose whether to go forward with the service and, in the event of a denial, to accept financial responsibility. If a claim is denied without the issuance of an ABN, then the therapist is not allowed to charge the beneficiary and is financially liable for the cost of any services rendered (CMS, 2013,

2018). Medicare is required to issue an initial determination of denial within 30 days of receipt of the claim, though they often reach this decision much faster (CFR, 2009).

Figure 1 summarizes the actions and outcomes at the therapy cap for patients, providers, and Medicare. In evaluating the cap's screening properties, we consider two key decision points in a patient's PT care trajectory: the week in which they "attempt" to bypass the cap via an exception, and the week in which they "approach" the cap ahead of a potential attempt. As a patient approaches the cap, they make a decision with their provider of whether to continue care or not. Patients who approach the cap but stop are considered "deterred" by the cap. We interpret this deterrence as a response to the ordeal associated with the cap—this includes the costs of care, cost of documentation, and the uncertain reimbursement net of denials. Those who are not deterred then continue care, produce documentation, and attempt to bill Medicare for care above the cap. Medicare then decides whether to approve or deny these requests. Patients who attempt to exceed the cap but are stopped by Medicare are classified as "denied," while those who continue past the cap are classified as "approved." Thus, the only way that they can go past the cap would be to make an attempt and to be approved. Section 2.3 describes in further detail how each of these outcomes are identified in our data.

Figure 1: Diagram of Therapy Cap Actions and Outcomes



Notes: This figure illustrates patients', providers', and Medicare's actions and outcomes at the therapy cap as discussed in Section 2.2.

In sum, a patient can be stopped at the cap for two reasons: deterrence or denial. Making this distinction is important because they capture two different channels through which the cap operates. Through the deterrence channel, providers and patients decide whether to try to continue care after assessing their private costs and benefits. Through denials, Medicare plays an active role in deciding who is allowed past the cap. We empirically estimate the size of each channel and characterize their screening properties separately.

2.3 Data, Sample Definitions, and Outcome Definitions

Data Our main source of data is the Medicare 20% Carrier claims files. These data include all Part B office-based spending for a random 20 percent subset of traditional fee-for-service Medicare beneficiaries. At the line-level, the key variables are procedure (HCPCS) codes, units, and final payments. Lines can be aggregated into claims, which each correspond to a single PT visit. At the claim-level, the key variables are diagnosis codes (ICD-9), billing modifier codes (including the KX modifier), dates of service, provider identifiers, and an indicator for payment denial. We define a "provider" as a unique combination of tax identification number (TIN) and state.[4] We use TINs to define providers as opposed to National Provider Identifiers (NPI) for two reasons. First, providers were only required to report NPI in 2008, which is after our analysis period. Second, we expect that many of the behaviors or investments providers would take up in response to the cap would be implemented at a *practice*-level, as opposed to at the individual provider level.[5] These include, for example, any upgrades to electronic medical record systems, improvements to documentation standards, or changes in screening practices.

We supplement the Carrier claims with additional spending and utilization measures from the 20% Medicare Provider Analysis and Review (MEDPAR) and 20% Outpatient files. We also use information on patient demographics, chronic conditions, prior utilization, and mortality from the Medicare Master Beneficiary Summary Files (MBSF). Finally, we use 2006 zip-level income statistics reported by the Internal Revenue Service (IRS, 2025).

Sample Definition We then restrict the sample to patients who receive in-office physical therapy.[6] We follow Amico et al. (2015) in identifying PT claims via the HCPCS code and the PT modifier code. We focus on the 91% of these patients who only see one provider for PT the entire year. Among these patients, we limit to patients with "regular" PT: those who have at least 5 weeks with at least $50 spending a week in the calendar year. We do this to eliminate outliers of patients who have short but expensive PT episodes (e.g., one week of over $1000 in spending with no spending in other weeks) due to concerns that these reflect misreporting or that these patients are not comparable to the rest of the sample.

We then create four different samples for each of our analyses; we summarize the sample definitions broadly here and describe them in further detail in the respective sections. The bunching analysis in Section 3.1 restricts to patients with end of year spending within

[4] We count TINs that operate in separate states as separate providers in order to capture the notion of a physical practice. 98% of TINs operate only in one state.

[5] In 2008, the average TIN is our sample is associated with 2.5 NPIs.

[6] The cap technically applied to PT *and* speech therapy combined, but we remove the 4 percent of patients who ever receive speech therapy.

[−$800, $1600] of the cap, the health analysis in Section 3.2 restricts to patients with an injury diagnosis and no PT in the six months prior to their first session, and the screening analysis in Section 4 restricts to patients who "approach" and "attempt" the cap (as defined below). Table 1 reports summary statistics for each sample in 2005 and 2006. Columns 1 and 2 show the summary statistics for the bunching analysis sample, columns 3 and 4 report statistics for the health analysis sample, and columns 5-8 report statistics for the "approach" and "attempt" samples.

Table 1: Summary Statistics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Bunching Sample		Health Sample		Approach Sample		Attempt Sample	
	2005	2006	2005	2006	2005	2006	2005	2006
Demographics								
Age	73.0	73.1	72.3	72.4	73.3	73.5	73.2	73.4
White	0.88	0.88	0.90	0.90	0.87	0.86	0.87	0.86
Female	0.65	0.65	0.66	0.66	0.65	0.65	0.65	0.65
Urban	0.85	0.85	0.84	0.84	0.86	0.87	0.87	0.88
ZIP code income ($)	67,699	68,735	67,166	68,131	70,668	71,508	71,216	71,910
Part D Low Income Subsidy	0.15	0.15	0.15	0.14	0.17	0.18	0.17	0.19
Dual eligible	0.14	0.15	0.14	0.22	0.16	0.17	0.16	0.18
Prior utilization and spending								
Any hospital stay last 6m	0.22	0.23	0.24	0.22	0.26	0.27	0.27	0.28
PT-related surgery last 6m	0.04	0.05	0.12	0.11	0.05	0.05	0.05	0.05
Pain procedure last 6m	0.31	0.33	0.35	0.36	0.32	0.34	0.32	0.34
In-office spending last 6m ($)	2088	2249	2313	2320	2538	2771	2620	2912
Predicted 12-month PT spending ($)	.	.	1450	1473	1618	1703	1649	1760
PT utilization/spending								
Number of visits	12.65	12.73	11.73	10.74	23.03	20.63	25.17	23.66
Number of weeks of PT	7.68	7.85	7.08	6.60	13.42	12.36	14.62	14.17
Total PT spending ($)	1161	1114	1035	890	2426	2030	2688	2345
Average spending per visit ($)	105.17	102.73	99.33	95.74	127.17	119.54	128.69	118.89
Week start PT	22	21	26	26	14	14	14	14
Observations								
Number of beneficiaries	125142	128841	62063	71019	38209	35923	31082	25126
Number of providers	13222	13084	9242	9520	8928	8784	8164	7354

Notes: This table reports summary statistics for the analysis samples in 2005 and 2006. Columns 1 and 2 show the summary statistics for the bunching analysis sample in Section 3.1, columns 3 and 4 report statistics for the health analysis sample in Section 3.2, and columns 5-8 report statistics for the screening samples ("Approach" and "Attempt") in Section 4. The predicted 12-month PT spending measure is calculated only for the health analysis and screening samples. Spending measures are in terms of Medicare spending and do not include patient cost-sharing. Receipt of the Part D Low Income Subsidy is measured in 2006 as the program was not available in 2005. Note that the 2005 and 2006 analysis are not mutually exclusive, meaning one patient can appear in the sample for both years. Data: 20% Medicare Carrier claims, Master Beneficiary files, MEDPAR, and Outpatient files.

Overall, within each sample the 2005 and 2006 subsamples are relatively balanced on measures of demographics and prior utilization. While the total number of visits and amount of PT spending typically decreases from 2005 to 2006, the samples are fairly balanced in terms of the average per-visit PT spending. The number of patients in the bunching and health samples increases from 2005 to 2006, which follows a pre-reform trend of increasing numbers of patients receiving PT in this time period. Looking closer to the cap, there is a slight drop in the number of patients who approach the cap and a notable decrease in the number who make an attempt from 2005 to 2006.

Denials and Attempts Denials can be defined at the claim, attempt-week, and patient-level. We classify a *claim* as "denied" if the Carrier claim payment denial code indicates the claim was denied. Ninety-four percent of denied claims for PT in 2006 are associated with no payment, while many of the remaining claims largely consist of partial payments.

To identify "attempt weeks"—weeks in which a patient attempts to go over the cap—we first distinguish between two ways to summarize the spending associated with a claim: the *billed* and the *paid* amount. The paid amount reflects the payment the provider receives after the bill has been processed, net of denials. This is the final payment amount reported directly in the claims, and is also what the therapy cap applies to. The billed amount reflects the care that was actually provided and that the provider demanded payment for, prior to any denials. Since we do not observe the amount of payment demanded for denied line items, we construct it using the procedure code. Specifically, we impute the billed amount using that provider-patient pair's average per-unit payment on claims without a denial for that procedure code.[7]

"Attempt weeks" are then defined as weeks in which the cumulative paid amount up to the prior week was below the cap, and the billed amount in that week plus the prior week's cumulative paid amount is projected to go over the cap. We classify an *attempt week* as "denied" if at least one of its associated claims has a denial. Conversely, an attempt week is considered "approved" if none of the associated claims have a denial (and thus the patient continues past the cap).

Finally, we classify a *patient* as "denied" if they make an attempt but never successfully make it over the cap that year—that is, their cumulative paid amount at the end of the year is below the cap. Conversely, a patient is "approved" if they end the year with a cumulative paid amount above the cap.

[7] If that provider-patient pair has never successfully billed for that procedure code, we use the provider's average payment. If the provider has never successfully billed for that procedure code, we use the average payment for that code among providers with the same Medicare Administrative Contractor. This approach is similar to that of Dunn et al. (2024).

Weeks from Cap We also assign to each week how many "weeks from the cap" the patient is. We only consider weeks in which a patient received PT care.[8] When the "weeks from cap" measure is negative, it denotes how many additional weeks of care a patient would have to receive to reach (and pass) the cap. When "weeks from cap" is (weakly) positive, it denotes how many weeks ago the patient passed the cap.

The method for assigning "weeks from cap" differs depending on whether a patient's cumulative paid spending at the end of the year fell above or below the cap, and if they are ever observed making an attempt to go over the cap. For patients who end the year above the cap, we define the first week they attempt to bill above the cap as week 0, and all other weeks with positive PT spending are defined relative to that week. For patients who make an attempt but never get successfully get past the cap (thus ending the year below the cap), their last week is considered week -1, and all other weeks are defined relative to that. Finally, for patients who never make an attempt and end the year below the cap, we extrapolate based on their prior weekly spending to calculate how many weeks away the patient is from reaching the cap. Specifically, we calculate the patient's 5-week rolling average of PT spending, take the difference between the cap value and the cumulative amount billed that week, and divide by the maximum of the patient's 5-week average *or* the average weekly spending in the sample.[9] Figure B1 illustrates week-level denial rates by "weeks from cap," and shows a clear increase at week -1 in denials that begins in 2006.

Approaches After defining weeks from cap, we can then define the week in which the patient "approaches" simply as the week where the patient is -1 weeks away from the cap. For patients who make an attempt, this means that their approach week is just the week before their first attempt. For patients who never make an attempt, it is the point at which one more week of their usual care is extrapolated to take them over the cap.

Documentation While we cannot directly observe documentation in the claims data, we proxy for it by looking for whether the "KX modifier" code is present on a claim. As discussed in Section 2.2, this modifier code was introduced as part of the 2006 therapy cap and is used by providers to indicate the availability of documentation justifying the medical necessity of spending over the therapy cap. Table 3 shows that having documentation, as indicated by the presence of this modifier code on at least one claim in an attempt week, substantially increases the likelihood that that week's attempt is approved. This correlation

[8] The week-to-week visit patterns in the data suggest that care is scheduled on a weekly basis. Figure B2 shows that PT care tends to be scheduled on the same day every week, as well as on a Monday/Wednesday/Friday or Tuesday/Thursday cadence.

[9] We divide by the maximum of the two to ensure that patients with low prior weekly spending are never implausibly far away from the cap (e.g., over 52 weeks away from the cap). Dividing by the maximum of the two effectively left-censors the "weeks from cap" measure at -8.

is robust to the inclusion of controls for patient and provider characteristics. Relative to a model with just patient demographics, health, and provider size, the R^2 increases 2-4-fold once this indicator is included in the regression.

Predicted PT Spending We also use the claims data to construct a measure of *ex ante* patient clinical need for PT. Using data from pre-reform years, we predict what a patient's 12-month PT spending would be absent the therapy cap, given their spending and utilization patterns prior to their first PT. We implement this using gradient-boosted decision trees from the LightGBM package. The predictors are age, sex, utilization and spending in the previous calendar year available in the MBSF Cost and Utilization file (in-office spending, Part B drug, outpatient procedure, inpatient, testing, imaging, hospice, evaluation and management, durable medical equipment, dialysis, and other), chronic conditions at the end of the previous calendar year, PT and OT spending in the previous calendar year, inpatient and SNF stays within the last 6 months (spending, number of visits, Diagnosis Related Group of their most recent inpatient stay, length of stay of their most recent inpatient stay, and days since last visit), in-office spending in the last six months, and an indicator for having an auto exception diagnosis in the last 6 months.[10] Importantly, note that the model is *not* trained on factors we later consider in our test for horizontal inequity: race and income (as well as zip code, which could be a proxy for both). The model is trained on patients who approach the cap in 2004 and 2005, prior to the implementation of the therapy cap. Thus, we are predicting what a patient *would* have spent on PT, absent the cap. We then apply this prediction model to patients who approach the cap in 2005 and 2006. We discuss the machine learning methodology and model fit in further detail in Appendix Section D.

3 Overall Effects of the Cap

3.1 Savings

Methodology and Sample Construction To quantify the Medicare savings from the cap, we apply methods from the "bunching" literature (Kleven, 2016). In our context, the pre-policy distribution serves as a natural counterfactual. In particular, we will compare distribution of Medicare PT spending in 2006 to the pre-reform distribution in 2005, restricting to the area around the 2006 cap—the "manipulation region." Medicare pays for 80 percent

[10] An "auto exception diagnosis" is one that CMS designated as not needing a written request to continue care past the cap "when services related to these conditions and complexities are appropriate provided and documented" (CMS, 2006b). In practice, we found that conditional on attempting to go over the cap, patients with a recent auto exception diagnosis faced similar denial rates (21%) as those without these diagnoses (25%), suggesting that this policy was not well-enforced in 2006. One reason for this could be that the auto exceptions policy was announced after the therapy cap was already in effect, as part of an update to implementation instructions that occurred in February 2006.

of allowed charges and patients are responsible for the remaining 20 percent, so the cap appears at $0.8 \times \$1740 = \$1392$ in the distribution of per-patient Medicare PT spending. We restrict to a region from \$800 below to \$1600 over this amount; this region has been chosen such that the "missing mass" to the left of the cap is approximately equal to the "excess mass" to the right. We interpret the difference in spending between the two distributions as the savings implied by the cap.

Rather than directly comparing the 2005 and 2006 distributions, we adjust the 2005 spending calculation to account for changes in procedure prices and in the total number of patients receiving PT from 2005 to 2006. Let $\bar{r}(F_j, p_k)$ be the per-patient spending in the manipulation region using the distribution in year j and prices in year k:

$$\bar{r}(F_j, p_k) := \int (q \cdot p_k)\, dF_j(q),$$

where F_j is a distribution in the region around the cap in year j over quantity q for each procedure, p_k is a vector of prices for each procedure in year k, and the integral is taken over all PT procedure codes. Thus, $\bar{r}(F_{06}, p_{06})$ denotes the actual average spending around the cap in 2006, and $\bar{r}(F_{05}, p_{06})$ denotes the average spending under the 2005 distribution, price-adjusted for 2006.[11]

In order to convert the difference between these two averages into a measure of total savings, we multiply the per-patient spending by the number of patients in this region in 2006, denoted as N_{06}. This ensures that the savings calculation is not affected by secular changes in the total number of patients receiving PT in this period. Thus, the savings in 2006 dollars and as a percent of 2005 spending are respectively:

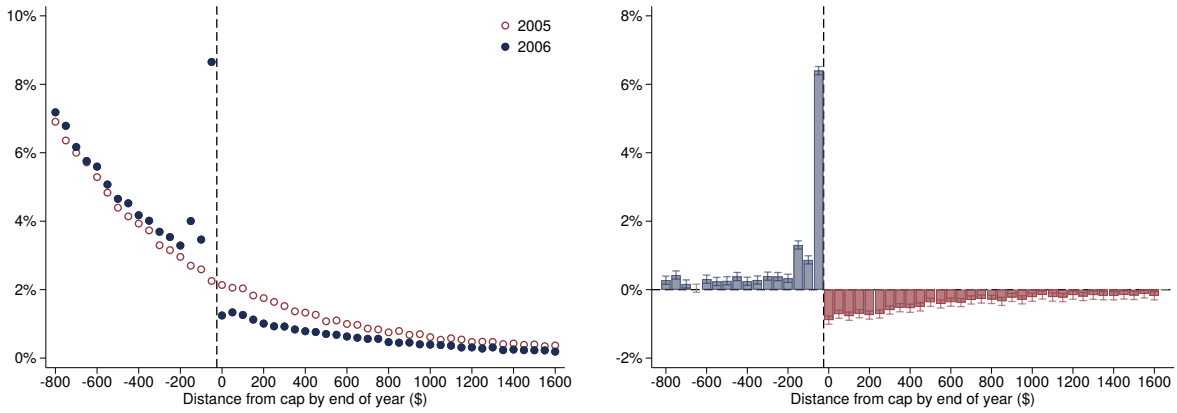
$$S_{06} := N_{06}\left(\bar{r}(F_{05}, p_{06}) - \bar{r}(F_{06}, p_{06})\right); \quad S^{\%}_{06} := \frac{\bar{r}(F_{05}, p_{06}) - \bar{r}(F_{06}, p_{06})}{\bar{r}(F_{05}, p_{06})}.$$

Results Figure 2 depicts the 2005 and 2006 spending distributions and their difference in the range from \$800 below the cap to \$1600 over the cap. The sum of the difference between the two distributions is \$83 million, or 7.6 percent of 2005 spending in the region around the cap. We estimate that there were 532,000 additional denials in this region in 2006, which implies that the return to Medicare per additional denial due to the cap was \$156.

[11] To impute prices, we use the provider-level 2006 price to account for geographic adjustments or any other provider-specific idiosyncrasies that affect the price they receive per procedure. For procedures that a provider rendered in 2005 but not 2006, we replace the 2005 price with the average 2006 price for that procedure in the same Medicare Administrative Contractor (MAC) region to account for any geographic differences in Medicare prices (League, 2023).

Figure 2: Distributions of Spending Around Cap in 2005 and 2006

(a) Distribution of Spending **(b) Difference between 2005 and 2006**



Notes: This figure plots (a) the distributions of end-of-year physical therapy spending around the cap in 2005 and 2006 and (b) the difference in the distributions between 2005 to 2006. Distance from cap is calculated in bins of $50 relative to the 2006 cap and shares are calculated as the share of patients within [-$800, $1600] of the cap. Data: 20% Medicare Carrier claims.

Interpreting the difference between the 2006 and 2005 distributions as the savings from the cap requires two key assumptions. The first is that the 2005 distribution captures what the 2006 distribution would have been absent the reform. This would be violated if this part of the distribution is not stable from year to year, in the absence of changes in Medicare policy. We can validate this assumption by comparing the 2005 distribution to the 2004 distribution, neither of which was subject to the therapy cap. Figure B3 shows that when comparing two consecutive years without a therapy cap, there is little difference in the spending distributions in this region.

The second assumption is that the introduction of the therapy cap did not change the share of patients who ended up inside or outside of the manipulation region around the cap. In other words, the only reason $N_{06} \neq N_{05}$ is due to secular trends over time in the total number of patients receiving PT and not a response to the therapy cap. This allows us to normalize patient count across the two years by multiplying by the 2006 patient count in this region. For this to be true, we need that there was no "extensive margin" response to the cap which differentially drew patients into or out of the manipulation region. It would be violated if, for example, providers who cut back on care for patients above the cap now have more capacity and use this additional capacity to accept more patients who end up far below the cap. In that case, the relative share of patients outside of the manipulation region would increase as a response to the cap.

We evaluate this second assumption in two ways. First, looking at the full distribution

of spending in 2005 and 2006 in Figure B4 panel (a), we note that the lower part of the distribution outside of the manipulation region is relatively stable. There does not appear to be any marked changes in the share of patients below the manipulation region. Second, if there were an extensive margin response, we would expect that it would be driven by providers who had relatively more patients over the cap in the pre-period, as they would be the ones experiencing the largest capacity expansions. In Figure B4 panel (b) we plot average patient count over time, splitting by providers who had high vs. low shares of patients over the cap in 2005. Providers with high 2005 over-cap shares do not appear to see more patients starting in 2006, indicating that they did not respond to the cap by taking on more patients at the lower end of the spending distribution.

Comparison to a Hard Cap To give the savings from the 2006 soft cap more context, we also compare it to the hard cap which was implemented in 1999. The difference between the two regimes is that there was no exceptions process in 1999 for patients to get care above the cap—effectively, all attempts to go past the cap were denied. The policy context for the 1999 cap is discussed in greater detail in Appendix Section C. Due to data limitations, we cannot compare 1999 to a pre-period year but instead compare it to 2000, after the cap was repealed. If there are any lingering effects of the 1999 cap in 2000, then this would bias our savings estimate downward.

Figure B5 plots the 1999 and 2000 spending distributions and differences in distributions in the range from \$700 (in 1999 dollars) below the cap to \$1300 above the cap, which is equal to approximately \$800 and \$1600 in 2006 dollars. Here, the savings are defined as:

$$S_{99} := N_{99} \left(\bar{r}(F_{00}, p_{99}) - \bar{r}(F_{99}, p_{99}) \right).$$

Taking the difference between the 1999 distribution and the price-adjusted 2000 counterfactual, we find that the hard cap reduced spending by 14.1 percent (\$23.4 million in 2006 dollars). Compared to the 7.6 percent reduction from the 2006 cap, this confirms that the 2006 therapy cap was indeed "soft"—the savings are diminished by almost half when providers are allowed to request exceptions to exceed the cap.

Quantifying the Deterrence and Denial Channels Figure 2 shows that the therapy cap causes bunching in the distribution of patient spending. To unpack what is driving this bunching, in Figure 3 we plot the distribution of end-of-year PT spending by the number of *weeks* a patient is from the cap. Panel (a) plots the overall difference between the 2005 and 2006 distributions by weeks to the cap. The excess mass to the left of the cap is concentrated in the last 3 weeks leading up to the cap, and there is a 93 percent (8.1 percentage point) increase in the share of patients who stop one week before the cap. This excess mass to the

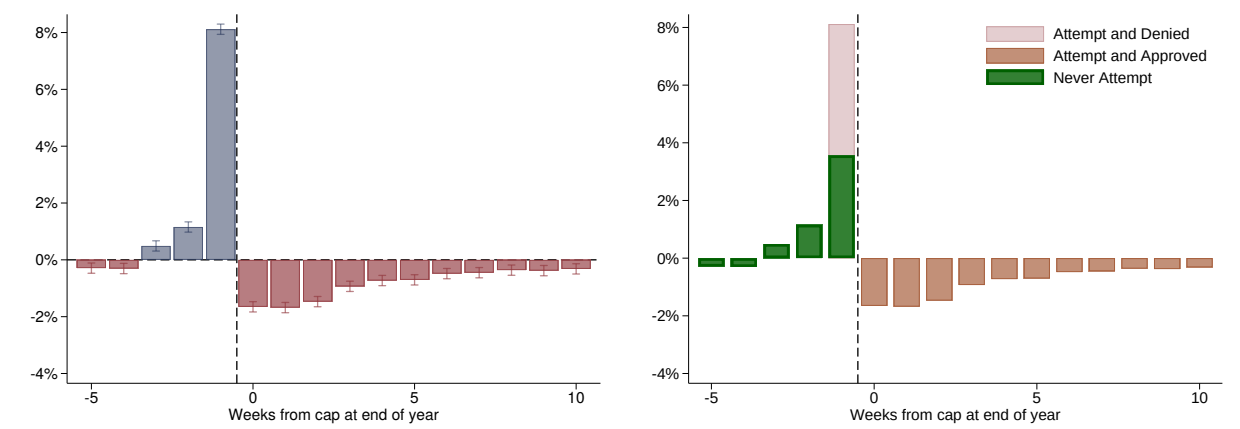
left of the cap comes from a reduction in the share of patients to the right of the cap, with effects extending even up to 10 weeks away from the cap.

Figure 3 panel (b) then decomposes the excess and missing mass into three categories. As in panel (a), we show the difference between the 2005 and 2006 distributions. Patients with an attempt to go over the cap can either be classified as "Attempt and Denied," meaning they made an attempt to bill past the cap but were denied and ended the year below the cap, or they are classified as "Attempt and Approved," meaning they ended the year above the cap. The remaining patients are classified as "Never Attempt," and end the year one or more weeks below the cap. Among patients to the left of the cap, we consider them to be "deterred" if they stop without an attempt, and "denied" if they attempt and are denied. Of patients who bunch in the week *immediately* before the cap, 42 percent stop due to deterrence while 58 percent stop because they are denied.

Figure 3: Distribution in Weeks to Cap, 2005-2006

(a) Difference between 2005 and 2006 distributions in weeks to cap

(b) Difference in weeks to cap, split by reason for stopping at cap



Notes: This figure plots the difference in the 2005 and 2006 distributions of cumulative physical therapy spending. Distance from the cap is calculated in terms of weeks of care relative to the 2006 cap and shares are calculated as the share of patients within [-5,10] weeks of the cap. Panel (a) shows the overall difference between 2005 and 2006, and panel (b) splits the difference between patients with an attempt to go past the cap in their last week but were denied ("Attempt and denied", red), patients with an attempt who were not denied ("Attempt and successful," orange), and patients with no attempt ("Never attempt", green). "Attempt and denied" is defined at the patient-level: it is an indicator for whether their cumulative paid amount at the end of the year is below the cap. Section 2.3 describes the construction of the weeks from cap measure. Data: 20% Medicare Carrier claims.

Comparing the distribution of beneficiaries allows us to decompose the share of beneficiaries who stop at the cap into the denial and deterrence channels. However, the decomposition of *savings* attributable to each channel could differ, as this depends not only on the number

of beneficiaries who stop, but also their counterfactual spending in the absence of the cap. This counterfactual spending is unobservable, but we can use the distribution of *predicted* spending to estimate the savings decomposition. Specifically, we use the predicted 2004-2005 PT spending measure described in Section 2.3 and Appendix Section D.

Figure B6 plots the distributions of predicted PT spending among "deterred" and "denied" patients who stop one week from the cap in 2005 (panel a) and 2006 (panel b). The savings attributed to each channel can be calculated as the difference between the area under the 2006 and 2005 distributions (panel c). We calculate that the savings from deterrence are $34 million and the savings from denial are $51 million, meaning 40 percent of the savings can be attributed to deterrence and 60 percent can be attributed to denial.[12] This split is very close to the share of beneficiaries who are deterred or denied, reflecting the fact that the spreads of the predicted savings distributions are fairly similar between the deterred and denied groups.

3.2 Patient Health Effects

We next consider whether the cap had a direct effect on patient health or led patients to substitute to other forms of care. If there is evidence of such effects, this would imply that the cap-induced savings were poorly targeted. To estimate the causal effect of the therapy cap on health outcomes, we use an instrumental variables (IV) strategy that leverages within-year variation in spending due to the differential "bite" of the cap, depending on when in the year a patient initiates PT. Appendix Section F presents an alternative approach which employs a difference-in-differences estimator and finds similar results.

Methodology and Sample Construction Our identification strategy leverages the fact that the therapy cap resets at the beginning of each calendar year, regardless of when a patient begins receiving PT care. Therefore, the cap is more likely to bind for patients that start their care earlier in the year. This should generate a negative relationship between when a patient starts PT and their total 12-month PT spending, once the cap is in place in 2006. Our instrument is therefore the month-year in which a patient starts PT care.

We construct our sample by identifying patients that start PT during 2005 or 2006.[13] The key exclusion restriction required is that conditional on controls, when a patient starts PT affects health outcomes only by making the cap more or less likely to bind. To ensure we are not capturing patients strategically timing when they start PT, we restrict to patients

[12] While most of the excess mass appears exactly one week short of the cap, there is also some excess mass dispersed up to three weeks away from the cap. Once patients who stop up to three weeks away are included in the deterrence channel, then 53 percent can be attributed to deterrence while 47 percent stop due to a denial. Under this categorization, the savings from deterrence rise to $54 million, or 51% of the total savings.

[13] We consider a patient to be "starting" PT only if they have received no PT for at least six months.

who received an injury diagnosis in prior 90 days (Appendix E). We also restrict to patients with 12-month PT spending over $200 to focus on patients more likely to be affected by the therapy cap.[14] As robustness checks, we re-run our analyses on low-income patients and high-need patients, who may be particularly vulnerable to spending reductions from the therapy cap. We also look for heterogeneous effects by decile of predicted patient need and by whether a patient had a recent pain procedure or orthopedic surgery.

We first estimate the following reduced form specification:

$$Y_i = \sum_{f=1}^{11} \theta_f \; 1(Year_{y(i)} = 2006) \times 1(FirstMonth_{m(i)} = f) + FirstMonth_{m(i)} + Year_{y(i)} + \varepsilon_i, \tag{1}$$

where θ_f is the coefficient on the interaction of an indicator for 2006 and the first month of a patient's PT, and the omitted month is December. $Year_{y(i)}$ is a year indicator and $FirstMonth_{m(i)}$ is an indicator for the patient's first month. Then the second stage is:

$$Y_i = \beta \widehat{PT_i} + FirstMonth_{m(i)} + Year_{y(i)} + \nu_i, \tag{2}$$

where $\widehat{PT_i}$ is instrumented 12-month PT spending based on equation (1), scaled so that β can be interpreted as the effect of an additional $100 in Medicare PT spending on the likelihood of each outcome. Standard errors are clustered by year and start month.

As our health outcome variables, we consider six indicators of patient health and utilization that could be related to an insufficient amount of PT: opioid prescriptions, pain management procedures, orthopedic surgeries, emergency department visits, inpatient stays, and skilled nursing stays. Given that the most common diagnosis for patients in our sample relates to muscle or joint pain, we might expect patients to substitute to opioids, pain management procedures, or orthopedic surgeries if PT did not successfully treat their pain. Patients also seek out PT to improve their strength and mobility—the top PT procedure code in our sample is for therapeutic exercises. Thus, an insufficient amount of PT could result in an injury or a fall that could result in an emergency department visit, an inpatient stay, or a skilled nursing stay.

Aside from opioid prescriptions, the outcomes are indicator variables that are measured within 12 months of a patient's last day of PT. We measure these outcomes starting after the last PT visit to capture patients seeking alternatives after PT has "failed," as opposed

[14] As discussed in Section 3.1 and Figure B4, we find no evidence of an extensive margin response to the cap in the lower part of the spending distribution.

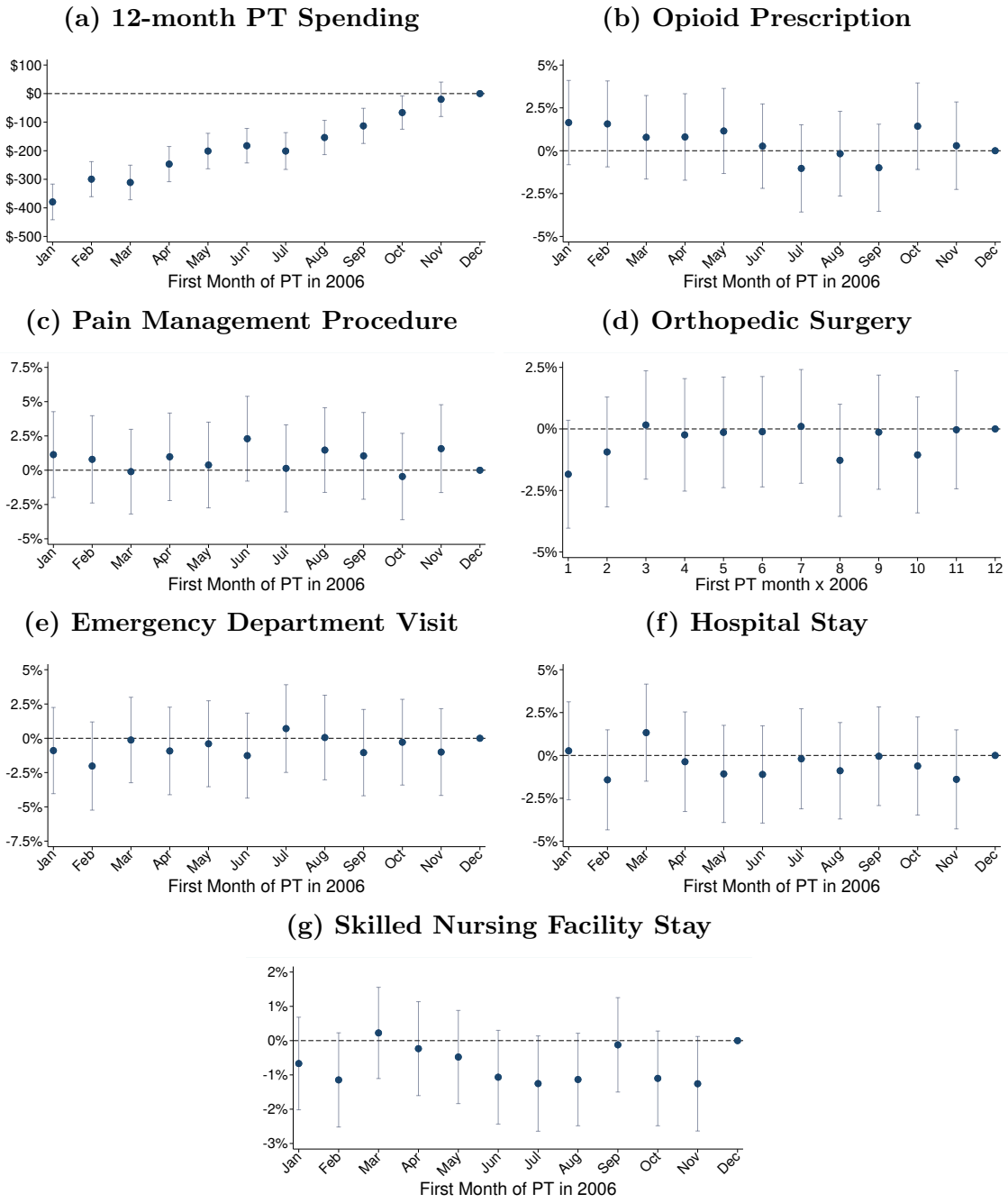
to contemporaneous utilization that could be endogenous to the PT's behavior—say, if a PT is co-located with a pain management practice. The opioid prescription outcome is an indicator variable that is measured between 12 and 24 months after the patient's last day of PT.[15] Appendix Section E provides further detail in how each outcome is constructed.

Before describing the results, it is important to consider what conclusions we can draw from our estimates, given our identification strategy and data. The therapy cap only restricts Medicare spending, so our estimates tell us the causal health effects of reducing *Medicare*-funded PT spending, rather than the effects of reducing *any* PT spending. If patients compensate for the cap by paying for the rest of their care out-of-pocket, then the therapy cap just represents a reduction in transfers, but not in actual PT utilization. We can address this partially by repeating our analyses just on low-income patients, for whom we might expect more passthrough of the cap savings onto total utilization. Furthermore, since we are using claims data, we cannot observe dimensions of health that PT is arguably most relevant for, such as mobility and pain levels. Our analysis is not be able to speak to whether patient well-being worsens in these dimensions unless the effects are large enough to induce patients to seek additional care.

Results Figure 4 shows the reduced form results from equation (1). Panel (a) plots the relationship between first month of PT, interacted with an indicator for 2006, and PT spending in the subsequent 12 months. Consistent with the cap binding more for patients who start earlier in the year, there is a monotonic negative relationship between start month and 12-month PT spending—the reduction in PT spending due to the cap is much larger for patients who start earlier in the year relative to those who start later. Turning to our health outcomes, we would expect that if the cap-induced savings were poorly targeted, then there would similarly be a negative relationship between these outcomes and a patient's start month. Instead, this relationship is flat and all coefficients are statistically insignificant at the 5 percent level.

[15] We construct this measure using this time frame because the Part D prescription data for opioids is only available starting in 2006, so we cannot measure opioid prescriptions within 12 months for the 2005 sample.

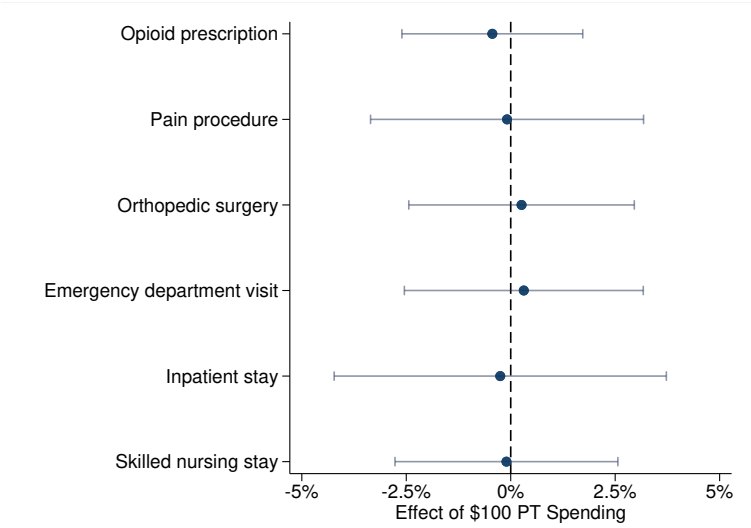
Figure 4: Reduced Form Spending and Health Outcomes



Notes: This figure plots the coefficient θ_f, which denotes the interaction between an indicator for 2006 and an indicator for month of first PT, from equation (1). Panel (a) plots the coefficients on 12-month PT spending ($). Panel (b) plots the coefficients on an indicator for opioid prescriptions, panel (c) plots the coefficients on an indicator for pain management procedures, panel (d) plots the coefficients on an indicator for orthopedic surgery, panel (e) plots the coefficients on an indicator for an emergency department visit, panel (f) plots the coefficients on an indicator for a hospital stay, and panel (g) plots the coefficients on an indicator for a skilled nursing facility stay. All outcomes other than opioid prescriptions are measured within 12 months of the first PT session; opioid prescriptions are measured 12-24 months after the first PT session. Section E describes how the health outcome measures are defined. Sample is restricted to beneficiaries with more than $200 in 12-month PT spending. Data: 20% Medicare Carrier, Outpatient, MEDPAR, and Part D claims.

In Figure 5, we plot the IV coefficients from estimating equation (2). Given our relatively strong first stage,[16] we are able to rule out out effect sizes larger than 4 percentage points in both directions for all outcomes.

Figure 5: IV: Effect of PT Spending on Health Outcomes



Notes: This figure plots the coefficient β, which denotes the effects of an additional $100 of PT on an indicator of each outcome, from equation (2). All outcomes other than opioid prescriptions are measured within 12 months of the first PT session; opioid prescriptions are measured 12-24 months after the first PT session. Section E describes how the health outcome measures are constructed. Sample is restricted to beneficiaries with more than $200 in 12-month PT spending. Data: 20% Medicare Carrier, Outpatient, MEDPAR, and Part D claims.

The lack of detectable health impacts on the overall sample could be masking effects for particularly vulnerable patients. We assess this possibility in three ways. First, we focus our attention on two patient populations where the reduction in Medicare PT spending could be particularly impactful. We first consider low-income patients, who may be unwilling or unable to pay for additional PT care out-of-pocket in the face of a Medicare denial. If so, the reductions in Medicare PT spending from the cap should pass through into relatively larger reductions in *overall* spending, which is confirmed by the larger reduced form effect on 12-month PT spending in Figures B7 and B8. Figure B9 shows that for two types of low-income patients—dual-eligibles and patients receiving the Part D LIS—there is no evidence that the reduction in PT spending due to the cap worsened health outcomes.[17]

Second, we stratify by a patient's initial health status, which is defined as their decile

[16]The Kleibergen-Paap Wald F-statistic for the first stage is 251.4.

[17]While the reduced form results show a statistically significant and positive estimate on opioid prescriptions for patients with the earliest start months, the IV estimate in Figure B9 for opioids is not statistically significant.

of predicted 2004-2005 PT spending constructed in Section 2.3. We estimate the IV in Equation 2 separately for each decile, where higher deciles indicate a patient is predicted to require more PT. We would expect to see negative effects to be concentrated mostly on relatively high-need patients. However, Figure B10 shows that even among the highest deciles of predicted PT spending, we find null effects on health outcomes.

Third, we stratify by whether the patient recently received PT-related care prior to starting PT. As shown in Table 1, about five percent of patients had an orthopedic surgery and about a third received a pain management procedure *prior* to starting PT. Thus, for these patients we may not expect that they would receive a second procedure or surgery after PT. Figure B11 stratifies the IV results by whether a patient reported having a pain procedure or orthopedic surgery within the last 6 months. Even among patients without a recent procedure or surgery, there is no evidence that reductions in PT spending led to increases in utilization of PT alternatives.

4 Characterizing Who is Screened Out by the Cap

Our results in the previous section indicate that the therapy cap saves money by screening out some—but not all—patients who approach the cap. The lack of a detectable health effect suggests that the cap screened out medically unnecessary care. In this section, we further explore *who* the cap screens out. We look both at dimensions which the cap was intended to screen on—the medical necessity of further PT—as well as dimensions it should *not* have screened on, like provider administrative capacity. Importantly, to understand the respective roles played by Medicare and by patients and providers, we characterize *how* the cap screens by studying how the denial and deterrence channels work both separately and together.

4.1 Screening on Need

We first ask whether the cap screened on medical need for PT—i.e., whether lower-need patients were more likely than higher-need ones to be stopped by the cap. We test for screening on need by considering how the share of patients denied, deterred, or stopped (for either reason) by the cap varies with patient need. If patients with higher need are less likely to be screened out, these relationships should be negatively-sloped. Of course, there could be many factors unrelated to the cap that could generate such a negative relationship—for example, if providers at baseline tend to give more care to higher-need patients, or if Medicare had other policies in place to deny unnecessary claims. Thus, to isolate the screening properties of the therapy cap specifically, we test for whether the slope of this relationship becomes *more negative* once the cap is in place.

Panels (a) and (b) of Figure B12 illustrate the intuition of this test, using the deterrence rate as an example. Panel (a) depicts a hypothetical case in which the cap deters care, but does *not* screen on need through deterrence. In particular, even though the deterrence rate shifts up, this increase in deterrence is the same regardless of need, leaving the slope of the relationship between deterrence and need unchanged. As a result, the marginal deterred patient has similar need as the average patient approaching the cap. In contrast, panel (b) depicts a hypothetical case in which the cap *does* deter based on medical need. In this case, the cap increases deterrence rates much more for low-need patients than high-need ones, resulting in a steeper relationship between deterrence and medical need. This is consistent with screening on need: the marginal deterred patient is relatively low-need. In Appendix Section G, we show that under a linearity assumption, this reduced-form slope test maps formally to the standard definition of screening within a potential outcomes framework: that the patients screened out by the cap—the compliers—are lower-need than those who are not.

We estimate the following regression to implement our screening test:

$$Y_i = \beta_1 X_i + \beta_2 X_i \times 1(Year_{w(i)} = 2006) + \gamma_{w(i)} + \varepsilon_i, \tag{3}$$

where Y_i is a dummy variable for deterrence, denial, or stopping at the cap due to either reason, X_i measures patient need, $\gamma_{w(i)}$ is a fixed effect for week-year w in which i approaches or attempts to exceed the cap, and $1(Year_{w(i)} = 2006)$ is an indicator for the patient attempting or approaching in 2006.[18] When the outcome is denials, the analysis is conducted on patients who attempt to exceed the cap. When the outcome is deterrence or stopping at the cap due to either reason, the analysis is conducted on all patients who approach the cap. β_1 captures the slope of the relationship in 2005, and β_2 captures how this slope changes in 2006. A negative, statistically significant β_2 indicates that the cap screened on medical need.

Our primary measure of patient need, X_i, is the predicted PT spending measure discussed in Section 2.3 and Appendix Section D. This measure is constructed by applying machine learning to claims from 2004-2005 and reflects the average amount of PT a patient would have been expected to receive in the years prior to the therapy cap based on their observable claims-based clinical history. Implicitly, this approach assumes that patients who tended to receive more PT care prior to the cap had greater underlying medical need, and a patient's rank in the predicted spending distribution is informative about their ranking in

[18] A given patient can spend multiple weeks where they are labeled as approaching or attempting the cap, and we include observations for each approach and attempt week to properly account for week-year time trends. Since all of a patient's approaches or attempts are with the same provider, we account for multiple observations per patient by clustering our standard errors at the provider-level.

the distribution of true medical need. We show in Appendix Section D that qualitatively, our findings still hold when simply using patient age as a proxy for medical need.

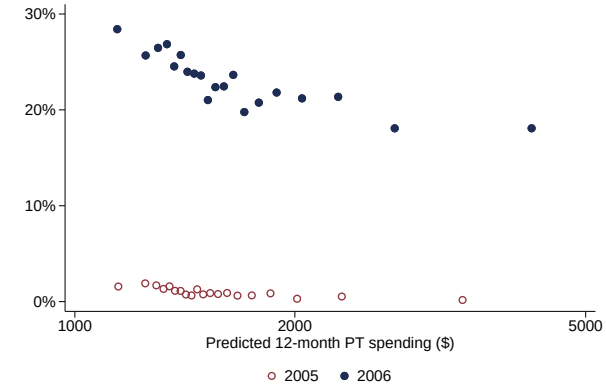
Results Figure 6 visually depicts our test for screening on need. Panel (a) plots binscatters of denial rates by patient need in 2005 and 2006, among patients who attempt to go over the cap value. Prior to the cap, Medicare was slightly more likely to deny claims at the cap threshold for low-need patients but, overall, denials were infrequent. Once the cap was introduced, Medicare became dramatically more likely to deny claims at this threshold: on average, the denial rate increased by 22.1 percentage points, a 2,100% increase relative to a pre-period average of 1 percent. Importantly, however, the magnitude of this increase was considerably larger for lower-need patients than higher-need patients. This steepened the relationship between denials and patient need, indicating that the cap screened on need through denials. Column (1) of Table 2 confirms this: the cap had a sizable and statistically significant negative effect on the slope of the relationship between denials and medical need.

Panel (b) plots binscatters of deterrence rates by patient need in 2005 and 2006, among patients who approach the cap. In 2005, the "deterrence" rate simply reflects the natural rate at which patients stop care at the cap threshold. Even in 2005, lower-need patients were more likely to stop at this level, consistent with screening on medical need occurring even at baseline. Once the cap was in place in 2006, deterrence increased by 7.9 percentage points (a 43% increase relative to the 2005 mean). Unlike denials, however, this increase in deterrence was uniform across the distribution of patient need. Correspondingly, the relationship between deterrence and patient need was no steeper after the cap was introduced than before. So while the cap generated a large deterrence response, it did not directly screen on need through deterrence. Column (3) of Table 2 confirms this, showing that the cap had a small and statistically insignificant effect on the slope of the relationship between deterrence and medical need.

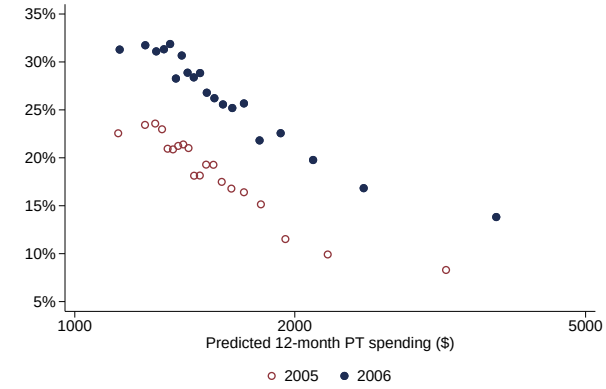
Our results indicate that the deterrence and denial channels have very different screening properties, reflecting differences in their baseline screening as well as in the effect of the cap. Prior to the cap, denial rates were low and only weakly related to patient need. Once the cap was introduced, denials increased substantially and the marginal denials were fairly targeted. In contrast, deterrence displays the opposite pattern. Even before the cap, lower-need patients were more likely to stop care at the cap threshold, reflecting substantial baseline "self-screening" in care decisions. The cap increases deterrence, but this marginal deterrence is largely uncorrelated with patient need. Thus, while the cap did deter care, there was little *screening* through deterrence.

Figure 6: Correlation between Patient Need and Denial and Deterrence Rates, 2005-2006

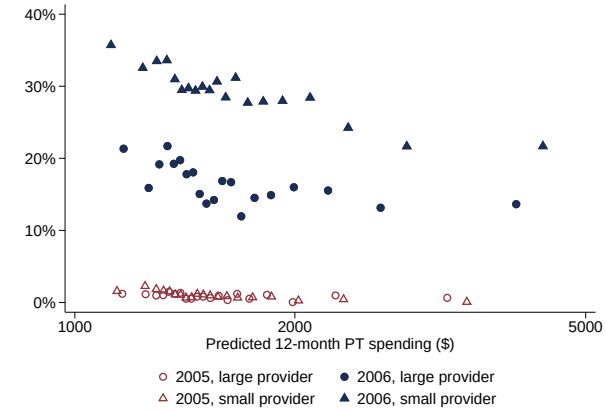
(a) Denied vs. patient need, conditional on attempt



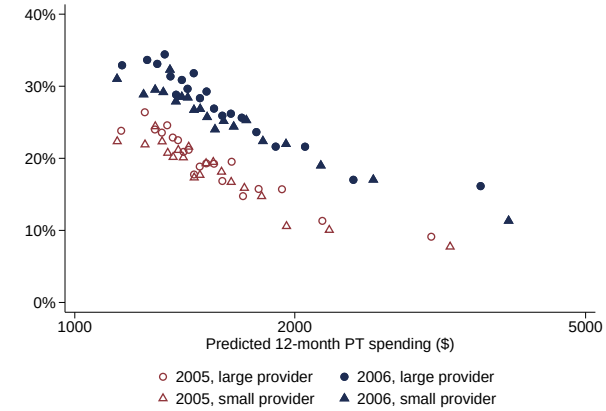
(b) Deterred vs. patient need, conditional on approach



(c) Denied vs. patient need by provider size, conditional on attempt



(d) Deterred vs. patient need by provider size, conditional on approach



Notes: This figure plots the relationship between cap outcomes and patient need and provider size. "Denied" is defined as the share of patients who attempt to go over the cap but are denied, meaning that their cumulative paid amount at the end of the year is below the cap. "Deterred" is defined as the share of patients who approach the cap but do not attempt. Panels (a) and (b) plot the relationship between log predicted 12-month PT spending and share denied and deterred in 2005 and 2006. Panels (c) and (d) plot the same relationships, split by provider size. Provider size is defined as the total number of Medicare beneficiaries who receive regular PT by that provider in 2006-2008, and a small provider denotes a provider with below-median patient count. Data: 20% Medicare Carrier claims and Master Beneficiary Summary File.

Table 2: Regression Results on Screening and Differences by Provider Size using Predicted Spending, 2005-2006

	(1)	(2)	(3)	(4)	(5)	(6)
	Denied | Attempt		Deterred | Approach		Stop at Cap | Approach	
Predicted Spending	-1.54***	-1.54***	-16.9***	-16.9***	-20.6***	-20.6***
	(.176)	(.177)	(.702)	(.703)	(.797)	(.798)
Predicted Spending × 2006	-6.22***	-6.86***	-.533	-.453	-3.29**	-3.72**
	(1.34)	(1.32)	(1.03)	(1.03)	(1.49)	(1.48)
Small provider		.154		-.741		-.371
		(.145)		(.513)		(.612)
Small provider × 2006		12.7***		-1.28		9.72***
		(1.06)		(.815)		(1.18)
Outcome mean, 2005	1.0	1.0	18.4	18.4	21.1	21.1
Outcome mean, 2006	23.1	23.1	26.3	26.3	52.2	52.2
Week-year FE	X	X	X	X	X	X
Cluster	Provider	Provider	Provider	Provider	Provider	Provider
N. Providers	11064	11064	11911	11911	11911	11911
N. Beneficiaries	53560	53560	70518	70518	70518	70518
N. Observations	116360	116360	80532	80532	80532	80532

Notes: * $p < .10$, ** $p < .05$,*** $p < .01$. This table presents the coefficients from estimating equation (3) with log predicted PT spending as X_i (columns (1), (3), and (5)) and equation (4) with log predicted PT spending as X_i and an indicator for whether a patient goes to a below-median-size provider as C_i (columns (2), (4), and (6)). Outcomes are all defined at the patient-level. All specifications include week-year fixed effects for the calendar week(s) that a beneficiary attempts or approaches the cap. The regression is clustered at the provider (TIN-state) level. Columns (1)-(2) restrict to the sample of 2005-2006 patients who ever make an attempt to go past the cap, while columns (3)-(6) restrict to the sample of who ever approach the cap, as defined in Section 2.3. Data: 20% Medicare Carrier claims.

The blunt deterrence response is particularly striking given the strong screening evident in the denial channel. The likelihood of a cap denial is much higher for low-need patients, meaning the expected reimbursement (i.e., reimbursement net of denials) is lower for these patients. If providers were responding solely to this difference in reimbursement, the deterrence response should also vary with need. Instead, the uniform deterrence response suggests that the decision to attempt to go past the cap was instead driven by factors uncorrelated with patient need. It could, for example, reflect the provider's opportunity cost of time spent on the visit, their willingness to bill patients in the event of a denial, or the patient's own ability to pay out of pocket if denied.

These contrasting channel-level screening effects naturally raise the question of how deterrence and denials come together to determine the overall screening of the cap. Because

the share of patients who stop at the cap is a non-linear function of the deterrence and denial rates,[19] the overall screening of the cap depends not only on its effects on the screening of each channel, but also on their levels. Specifically, screening through one channel is weakened when the level of the other channel rises. To see this, note that because deterrence increased in general, fewer patients reached the point where denials could apply, and so the impact of screening on denials is attenuated. Put differently, even strong screening on denials may not matter if only a small share of patients ever reach the point where they could face a denial. Thus, we separately test for the overall screening of the cap by estimating equation (3) using stopping at the cap due to deterrence *or* denial as the outcome variable. Column (5) of Table 2 shows that the slope of the overall stopping rate with respect to need becomes more negative in 2006. Thus, despite the increase in deterrence, the screening on need through the denial channel dominates, and overall the cap screens out relatively low-need patients.

4.2 Horizontal Inequity: Screening on Other Characteristics

Our second set of analyses then examines horizontal inequity conditional on need—whether patients with the same level of need face different probabilities of being screened out by the cap in a systematic way. Because the cap introduced a documentation requirement, we pay particular attention to variation related to provider administrative capacity, which we proxy for using provider size. Furthermore, since patients sort non-randomly across providers, we further explore whether differences across providers translate into disparities across patient groups.

We test for cap-induced *level differences* across groups defined by a characteristic C_i (e.g., whether a patient goes to a small provider). Specifically, we say that horizontal inequity arises if, holding need fixed, outcomes diverge between patients with $C_i = 0$ and $C_i = 1$ once the cap is introduced. Figure B12 panel (c) illustrates this intuition: while the deterrence rate is similar across groups in the pre-period, a gap emerges in the post-period, with patients in the $C_i = 1$ group more likely to be deterred at the same level of need. We implement this test by estimating the following equation:

$$\begin{aligned} Y_i = \beta_1 X_i + \beta_2 X_i \times 1(Year_{w(i)} = 2006) + \beta_3 C_i \\ + \beta_4 C_i \times 1(Year_{w(i)} = 2006) + \gamma_{w(i)} + \nu_i, \end{aligned} \tag{4}$$

where C_i is an indicator variable for whether a patient goes to a small provider or a patient demographic characteristic. β_3 captures whether patients with $C_i = 1$ are at baseline more

[19] Specifically, $S_i = D_i + (1 - D_i)Q_i$, where S_i is an indicator for stopping at the cap, D_i is an indicator for deterrence, and Q_i is an indicator for being denied conditional on attempt.

or less likely to be deterred or denied, and β_4 captures whether this changes once the cap is in place. Thus, our test for whether the cap introduced horizontal inequity is whether β_4 is statistically significant. Importantly, these must be differences that arise after controlling for patient need, X_i.

By Provider Size Given that the cap introduced a documentation requirement, we first consider differences by provider administrative capacity. We proxy for administrative capacity using provider size, as prior work has shown that larger providers tend to have an advantage in billing and documentation (League, 2023; Dunn et al., 2024). Differences by provider size could arise in the deterrence stage if larger providers have lower documentation costs and are across the board more likely to make attempts, or could manifest in the denial stage if larger providers have greater awareness of billing rules or better compliance with paperwork requirements.

Figure 6 panels (c) and (d) plot the same correlations as before between patient need and denials or deterrence, but split by whether a patient goes to a provider with above-median Medicare patient count—a "large provider"—or below-median—a "small provider."[20] Panel (c) shows that once the cap is introduced, patients who go to small providers are much more likely to be denied. The gap in denials between large and small providers exists conditional on patient need, meaning it cannot be attributed to differences in patient composition across providers. It instead must be driven by differences in *provider* behavior that influence the denial rate, independent of patient characteristics. The estimates from equation (4) in Table 2 column (2) confirm this formally: conditional on need, the gap between large and small providers grows by 12.7 percentage points in 2006—over half of the average 2006 denial rate of 23 percent.

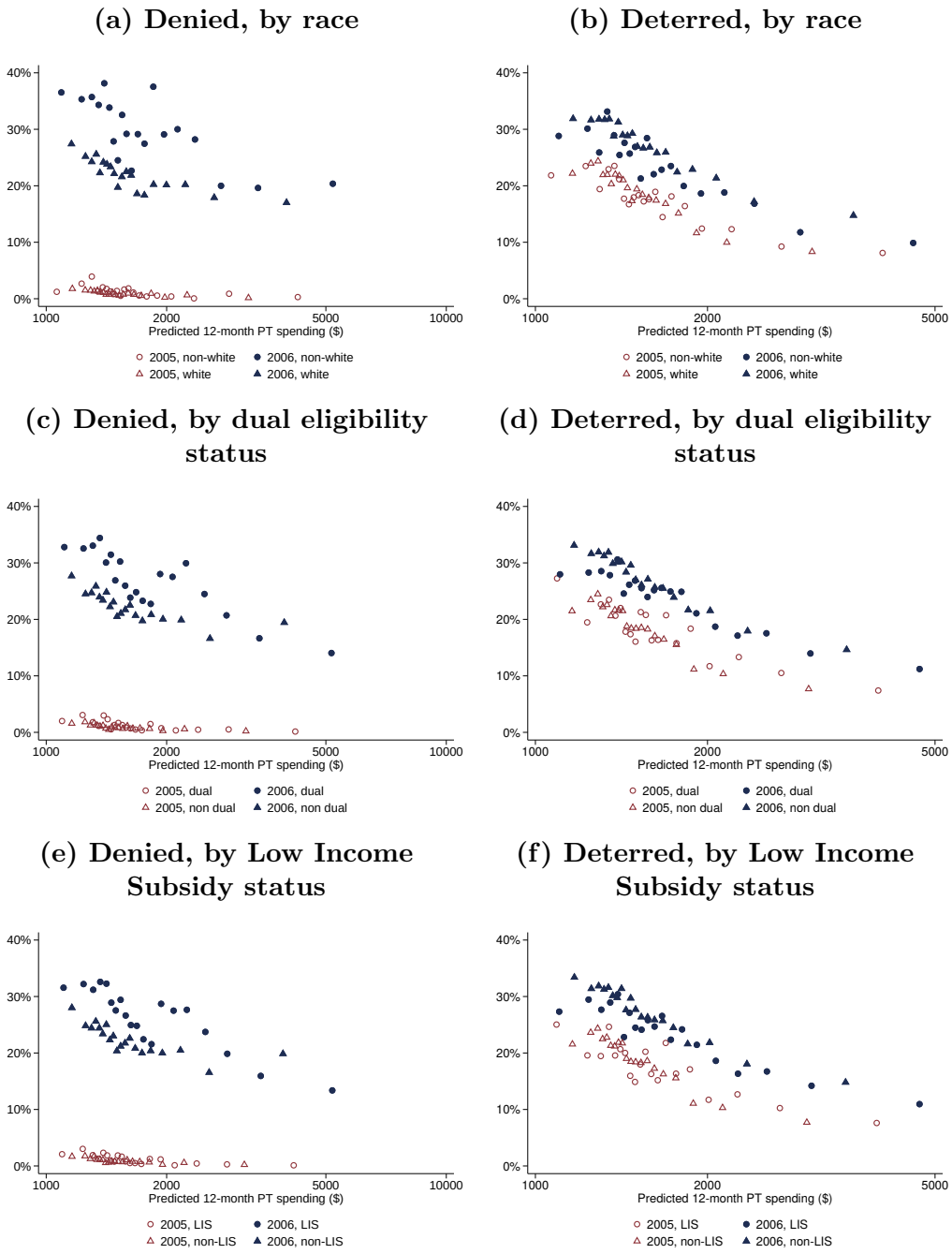
In contrast, when we turn to the deterrence response in panel (d), there appears to be no difference in deterrence across large or small providers in either 2005 or 2006. This is reflected in the statistically insignificant coefficient on "Small provider $\times$ 2006" in Table 2 column (4). Given the difference in denial rates by provider size, the lack of a similar difference in deterrence suggests that smaller providers may face information frictions about the therapy cap and the associated denials. We explore the drivers of these potential frictions in Section 5. Finally, in Table 2 column (6), we show that the higher denial rates for small providers translate into a higher share of their patients being stopped by the cap. Holding fixed medical need, patients of small providers are 9.72 percentage points (19 percent) more likely to be stopped than their counterparts at large providers in 2006.

[20]The median patient count in our sample is 925 Medicare patients in total between 2006 and 2008.

By Patient Race and Income The difference in denials between large and small providers is itself noteworthy because it suggests that the cap is screening on a characteristic unrelated to patient need, in contrast to the intention of the policy. However, these differences would have limited implications for horizontal inequity across *patients* if patient sorting across providers is largely random. But if different patient groups systematically see different-sized providers, then screening on provider size is more troubling. Indeed, we find strong evidence of non-random patient sorting: lower-income and minority patients tend to go to small providers: they come from lower-income zip codes, are more likely to be dually-eligible for Medicaid and receive the Part D Low Income Subsidy (LIS), and are more likely to be non-white (Figure B13).

These patterns motivate the analyses in Figure 7, which repeats the horizontal inequity analysis in Figure 6, splitting by patient demographic groups. Table A3 shows the results from estimating Equation 4 using patient demographics as C_i. Columns (1)-(3) show that the gap in denials across large and small firms translates into differences in denials by patient race and income: among those who attempt, non-white and low-income patients are 20-38 percent more likely to be denied by the cap than their respective counterparts. Interestingly, on the deterrence margin, columns (4)-(6) show that lower-income and minority patients are actually slightly *less* likely to be deterred by the cap, as indicated by the negative coefficient on the interaction between each demographic characteristic and the 2006 year indicator. On net, the horizontal inequity in denials dominates—non-white and lower-income patients who approach the cap are 8-11 percent more likely to be stopped by it, conditional on patient need (columns (7)-(9)).

Figure 7: Correlation between Patient Need and Denial or Deterrence by Patient Demographics, 2005-2006



Notes: This figure plots the relationship between cap outcomes and patient need and demographics. "Denied" is defined as the share of patients who attempt to go over the cap but never successfully make it past the cap. "Deterred" is defined as the share of patients who approach the cap but do not attempt. Panels (a) and (b) plot the relationship between log predicted 12-month PT spending and the denial and deterrence rates in 2005 and 2006, split by patient race. Panels (c) and (d) plot the same relationships, split by patient dual eligibility status for Medicaid in the respective year. Panels (e) and (f) plot the same relationships, split by whether the patient receives the Part D Low Income Subsidy in 2006. Data: 20% Medicare Carrier claims and Master Beneficiary Summary File.

The emergence of the gap by provider size in Figure 6 and the strong correlation between provider size and patient demographics in Figure B13 are consistent with patient sorting across large vs. small providers as being the key driver of the race and income gaps in Figure 7. But an alternative explanation could be that these demographic gaps reflect across-group differences unrelated to the identity of a patient's provider. Differences in deterrence could arise from low-income and minority patients being less healthy in a way that is observable to PTs but not captured in our predicted spending measures. Differences in denials could be driven by Medicare's denial algorithm inheriting some bias embedded in the data-generating process (Obermeyer et al., 2019). Such across-group differences would generate demographic disparities both across- and *within*- providers. We test this alternative explanation by estimating equations (3) and (4) with the inclusion of provider-year fixed effects in Table A4. Once provider fixed effects are included, the disparities by income become statistically insignificant and the disparity by race shrinks substantially. This confirms that the disparities in Figure 7 are largely driven by patient sorting *across* different providers—low-income and minority patients tend to go to providers who are less-adept at navigating the cap.[21]

Taken together with the results on screening on need, these results demonstrate that "soft" screening mechanisms like the therapy cap introduce an equity-efficiency tradeoff. On the one hand, the cap screens out low-value care—patients with lower *ex ante* need see larger increases in denials than patients with higher need from the cap. On the other hand, the cap also introduces a substantial advantage for some providers. Prior to the cap, two patients with similar medical need but who see differently-sized providers received similar amounts of care; after the cap, the patient of the smaller provider would receive less care. This "screening" by provider size appears to be driven by a factor unrelated to patient need. In the next section, we argue that it is driven by across-provider differences in administrative capacity, specifically in their knowledge of and compliance with the documentation requirement.

5 Drivers of Provider Size Advantage

Section 4 demonstrates that there is substantial variation in denials across large and small providers, which translates into differences across patients that are orthogonal to their medical need. In this section, we show that these across-provider differences are driven by variation in compliance with the cap's documentation requirement. Furthermore, large providers tend to do better because they have more opportunities for learning-by-doing with the cap.

[21] Furthermore, the model which includes provider fixed effects shows that the improvements in Medicare screening on denials are *not* the result of patient sorting. This can be seen by the negative and statistically significant estimate for the "Predicted staffing × 2006" coefficient in Table A4, column (1).

5.1 The Role of Documentation

While documentation review was not conducted for every approved exception to the cap, CMS emphasized the importance of always having documentation *available* for review in its communications about the therapy cap (CMS, 2006c). Providers were instructed to add a modifier code (the "KX modifier") to their claims to attest that documentation was available. We interpret the use of the modifier code as an indicator of baseline provider awareness about the requirement to have documentation on hand to support their request.[22]

Documentation use is strongly associated with denials both in the cross-section and over time. Table 3 first explores the cross-sectional variation and reports the estimates of a regression between an indicator for using documentation on an attempt and the likelihood that the attempt is approved (i.e., not denied). There is a strong positive correlation between the two, and the inclusion of documentation approximately doubles the likelihood that an attempt is approved. The magnitude of the coefficient is stable even with the inclusion of patient demographics, patient predicted health, and provider size. Including the indicator for having documentation into the regression model increases the R^2 by 2-4-fold.

Table 3: Regression of Documentation on Approvals on Cap Attempts, 2006

	(1)	(2)	(3)	(4)
		Approved | Attempt		
Has documentation	22.41***	21.89***	22.49***	20.85***
	(0.887)	(0.905)	(0.880)	(0.908)
N	51049	43620	54455	43620
R^2, no modifier code	.018	.019	.024	.032
R^2, with modifier code	.068	.067	.074	.075
Outcome mean	46.88	47.02	46.25	47.02
Patient demographics	X	X		X
Patient health		X		X
Provider size			X	X

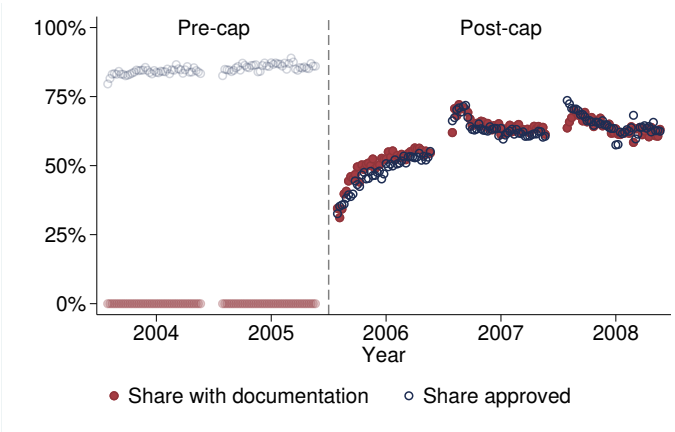
Notes: This table reports the relationship between approvals and documentation on attempts in 2006. "Approved | Attempt" is defined at the attempt-week level, and is the share of claims where the patient faces no denials in their attempt week. "Has documentation" is an indicator variable for whether any claim submitted that week has a KX modifier code. The table reports two R^2 values: one is for the regression which includes the "Has documentation" variable in addition to all the controls listed below, and the other is for a regression which includes only the control variables. The patient demographic control variables are age, race, and sex. Patient health controls are predicted 12-month PT spending, in-office spending in 6 months prior to first PT, total inpatient spending last year, SNF spending last year, total Part B spending last year, and total imaging spending last year. Provider size is measured as the total number of attempts the patient's provider made that year. All specifications include week-year fixed effects for the calendar week(s) that a beneficiary attempts to go over the cap. Standard errors (in parentheses) are clustered at the provider level. Data: 20% Medicare Carrier claims and Master Beneficiary Summary File.

[22] We cannot directly observe the documentation itself or verify that all providers who claimed to have it actually did.

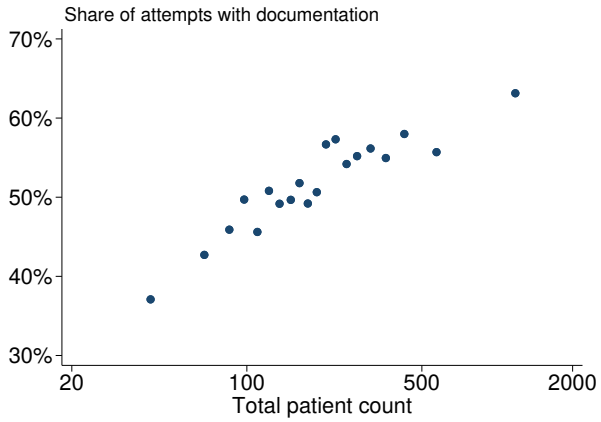
We next examine the variation in documentation use and denials over time. Figure 8 panel (a) plots the share of attempts with documentation and the share of attempts approved between 2004 and 2009. Prior to the introduction of the cap, documentation use is mechanically zero, since the modifier code was not introduced yet. Strikingly, once the cap is in place, the claim approval rate moves in lock-step with the use of documentation. This co-movement strongly suggests that changes in documentation use are driving the increase in approvals over time.[23]

Figure 8: Documentation, Approvals, and Provider Size

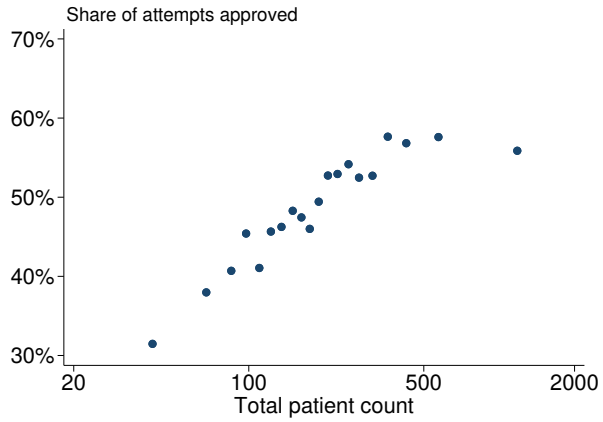
(a) Time Series, 2006-2008



(b) Documentation vs. Provider Size



(c) Approvals vs. Provider Size



Notes: Panel (a) plots the share of attempts with at least one claim with documentation ("Share with documentation") and without any denials ("Share approved") in 2006-2008. Approval is defined at the attempt-week level and is the share of claims where the patient faces no denials in the week that they make an attempt, where the attempt week is defined as in Section 2.3. "Share with documentation" is the share of attempts with at least one claim using the KX modifier code. Panels (b) and (c) plot the relationship between provider size and approvals and documentation use (respectively) on attempts in 2006.. Provider size is measured as a TIN-state's 2006-2008 Medicare patient count. Data: 20% Medicare Carrier claims.

[23] In Appendix Section D, we use a machine-learning model to rule out strategic *billing* changes—upcoding— as an explanation for the change in approval rates over time.

Documentation use (and therefore the approval rate on attempts) is also highly correlated with provider size: large providers are much more likely to use documentation on attempts than small providers and much more likely to be approved (Figure 8 panels (b) and (c)). Providers in the top quartile of size, who see about 1500 Medicare patients from 2006-2008, are 22 percent (11.8 percentage points) more likely to use documentation on their attempts in 2006 than those in the bottom quartile, who see 575 patients. This translates into a sizable advantage in approvals: their attempts are 48 percent (16.1 percentage points) more likely to be approved than those in the bottom quartile.

5.2 Link between Size and Compliance: Learning-By-Doing

Decomposition of Provider Size Advantage We next explore *why* larger providers have better compliance with the documentation requirement and thus fare better in the face of the cap. Documentation practices could vary across providers of different sizes for many reasons. On the one hand, large providers may just differ at baseline: they could have made larger initial investments in technology like electronic health records or have more knowledgeable billing staff. The pattern could also reflect an omitted factor that is correlated with both provider size and documentation use, like ownership status or geographic location.

On the other hand, the increase in documentation and approvals throughout the first year in Figure 8 suggests an alternative mechanism for the size advantage: learning. If providers learn through their experience with the cap—"learning-by-doing"—then large providers naturally derive an advantage from the ability to move up the learning curve more quickly. Figure 9 panel (a) shows that small providers accumulate experience with the cap much more slowly than large providers. By the end of the first year, approximately 60 percent of small providers still have less than 10 patients worth of experience with the cap, whereas the majority of large providers have more than 20 patients worth of experience.

Motivated by this distinction, we next decompose the variation in documentation use into learning-by-doing and persistent size-based advantages. We estimate a linear model of an indicator for documentation use in the week of an attempt, Y_i, for patient i receiving care from provider j that is attempting to bill above the cap in week t:

$$Y_i = \underbrace{\sum_{e(j(i),t(i))=1}^{50} \kappa^e}_{\text{Learning-By-Doing}} + \underbrace{Week_{t(i)}}_{\text{Industry-Wide Trends}} + \underbrace{\alpha_{j(i)}}_{\text{Provider FE}} + \underbrace{\beta X_i}_{\text{Patient Controls}} + \varepsilon_i \tag{5}$$

The regression decomposes the variation into three key components of interest. The first is κ^e, which captures the average amount of learning-by-doing by a provider who has experienced

$e(j(i), t(i))$ previous patients approaching the cap. The second is a calendar week fixed effect $Week_{t(i)}$, which captures any industry-wide trends affecting all providers.[24] The last is $\alpha_{j(i)}$, a provider fixed effect that captures any persistent provider-level difference in documentation use. This captures baseline variation in provider aptitude for billing that may result from investments or other provider characteristics. We also control for X_i, the predicted need of the patient associated with the attempt.

Figure 9 panels (b)-(d) present the estimates from equation (5). Panel (b) plots κ^e for $e \in [1, 50]$; the estimates suggest a significant amount of learning-by-doing, especially earlier on. The implied advantage from being further up the learning curve is large: a veteran provider with more than 50 patients worth of experience having a 20 percentage point advantage over a complete novice. Panel (c) plots $Week_t$ and shows significant industry-wide increases in approvals in the first year of the policy. Finally, panel (d) plots a binscatter of provider fixed effects α_j against provider size. The lack of a relationship between the fixed effects and provider size indicates that, after accounting for differences in experience, larger providers do not have a fixed advantage in compliance with the documentation requirement. Taken together, the results indicate that the provider size advantage in documentation stems primarily from learning-by-doing, which mechanically accumulates faster for larger providers.

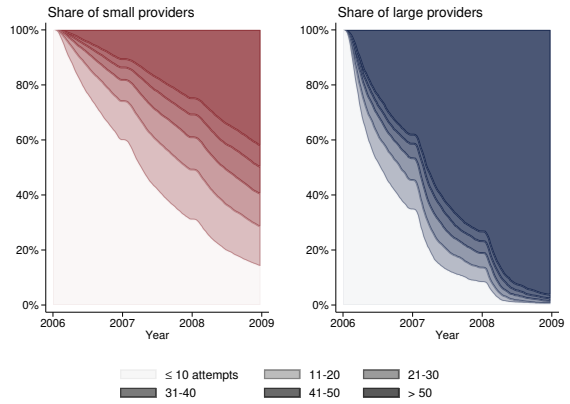
We repeat this decomposition analysis in Figure B14 where the outcome is approvals of attempts instead of documentation use. Here, we also find a learning curve consistent with learning-by-doing, but the provider fixed effects indicate that even after accounting for learning-by-doing about documentation use, large providers have some advantage on approvals. This suggests that larger providers may have some baseline advantage with approvals that extends beyond basic documentation use. For example, they could have access to technology to produce richer documentation of medical necessity or more knowledgeable billing staff.

Finally, in Section H, we complement our decomposition analyses with additional event study evidence on sharp within-provider learning. We look at how documentation use and approvals evolve around an event that should be associated with learning: the first time a provider successfully reverses a cap denial. Consistent with learning-by-doing, we find after the first time this happens, providers are consistently more likely to include documentation and be approved on all subsequent attempts.

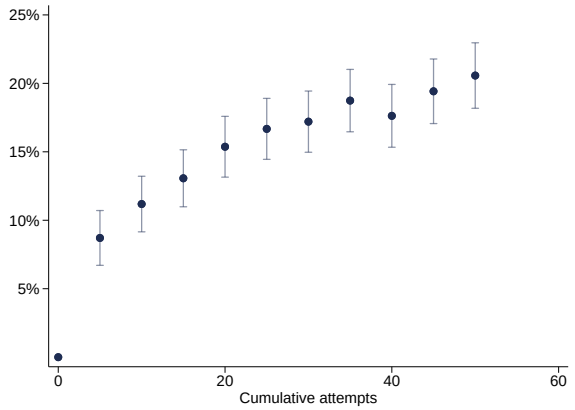
[24] This could capture, for example, widespread dissemination of information about documentation requirements across all providers by CMS or industry groups.

Figure 9: Decomposition of Size Advantage on Documentation Use

(a) Cumulative Experience, by Provider Size

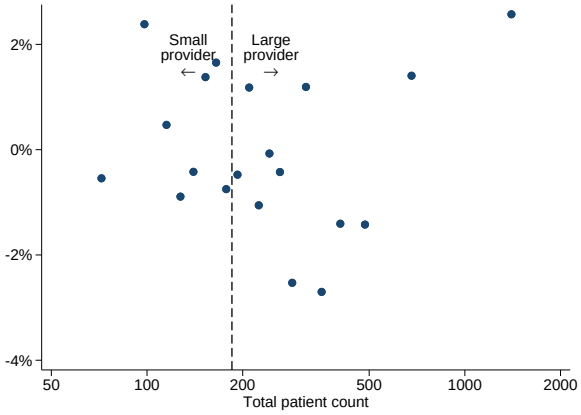


(b) Learning-by-Doing (κ^e)



(c) Industry-Wide Time Trend ($Week_t$)

(d) Provider Fixed Effects (α_j) vs. Size



Notes: This figure characterizes differences in documentation use on cap attempts by provider size and experience in 2006-2008 using data and regression estimates from equation (5). Panel (a) plots the share of large and small providers who achieve a cumulative number of cap attempts over time, where large providers are defined as having above-median patient count in 2006-2008. Panel (b) plots the coefficient on the provider's cumulative number of prior attempts on approvals. Approval is defined at the attempt-week level, and is the share of claims where the patient faces no denials in the week that they make an attempt, where the attempt week is defined as in Section 2.3. Panel (c) provides estimates of weekly industry-wide trends. Panel (d) plots the provider fixed effects against provider patient count. Section 2.3 describes the definition of cap attempts in further detail. Data: 20% Medicare Carrier claims.

Policy Implications We have shown that patients of large and small providers face substantially different approval rates in their attempts to go over the cap in the first year. Furthermore, our decomposition and event study results demonstrate that one key driver of this size advantage is learning-by-doing. This implies that, in the long-run, small providers' documentation and approval rates should approach that of large providers. Thus, while this

short-run horizontal inequity across providers is likely undesirable, the longer-run implications of the size advantage depend on whether the policymaker's desired approval rate is closer to that of small or large providers—that is, whether their eventual goal is for exceptions to the cap to be relatively rare or frequent.[25]

If the goal is to maximize savings and for cap approvals to be relatively rare—as they are for small providers—then the fact that there is rapid provider learning means that the cap's efficacy will diminish quickly over time. Consistent with this, Figure B16 shows that the excess and missing mass around the cap decrease substantially from 2006 to 2008, meaning the overall cap savings are diminishing over time. One way to mitigate this is to make Medicare's approval process less deterministic and thus more difficult to "learn" about. Instead of having approvals depend mostly on whether the provider simply indicates that they have documentation, approvals could be tied more directly to the content of the documentation. The tradeoff here would be that in-depth documentation reviews are more costly for both Medicare and providers.[26]

However, if the goal is for the cap to be fairly "soft," meaning approvals are relatively frequent—as they are for large providers—then in order to get ahead of horizontal inequity by size, Medicare will need interventions that level the playing field up front for smaller providers. Rather than letting providers learn through cumulative experience, Medicare could instead provide targeted provider education about billing rules or subsidize related technology, both of which are tactics it has used before in other contexts (CMS, 2014, 2024b).

6 Conclusion

This paper studies how soft spending limits screen by examining a spending limit imposed on healthcare providers by Medicare for physical therapy. We find that the soft "therapy cap" reduced overall spending, though much less than when compared to a hard cap that restricts all spending above the cap. We look for direct effects of these savings on patient health using an identification strategy that leverages the annual nature of the cap, which makes the cap more binding for patients who start PT earlier in the year. We find no evidence that cap-induced spending reductions resulted in increased use of PT substitutes like opioid use, pain procedures, and orthopedic surgeries, as well as of ED visits and inpatient or skilled

[25] In Figure B15, we re-run the health analysis from Section 3.2, split by patients of above- or below-median-sized providers. We find that there is no detectable health effect from cap-induced savings in either type. This suggests that making cap approvals relatively rare, as is the case for small providers, would not cut medically necessary spending.

[26] For example, Medicare's Recovery Audit Contractor program contracted with clinicians to conduct in-depth documentation reviews of medical necessity (Shi, 2024).

nursing stays.

Using a novel feature of healthcare claims data which allows us to observe both successful and unsuccessful attempts to bypass the cap, we differentiate between the deterrence channel—those who stop just short of the cap without making an attempt—and the denial channel—those who make an attempt but are not approved by Medicare. We then characterize which patients are stopped at the cap and whether they stop due to deterrence or denial. Both channels contribute to the overall savings, with denials contributing slightly more than half. However, the two channels differ in their screening properties. While denials screen on need, meaning lower-need patients were more likely to be denied than high-need ones, deterrence does not. This implies that the cap's overall screening on need is driven by Medicare's discretion in which cases to deny, rather than efficient "self-screening" by patients and providers.

Furthermore, we find that the cap introduces inequities which did not exist before: conditional on need, patients who go to small providers faced much higher denial rates and were therefore more likely to be stopped by the cap. As lower-income and minority patients tend to see smaller providers, these differences across providers translate into sizable disparities by race and income in the effect of the cap among otherwise-similar patients.

These differences appear to be driven by heterogeneous administrative capacity across providers. All else equal, attempts to go over the cap are much more likely if providers indicate they have documentation of need, and large providers are much more likely to do so. We decompose the size advantage into the components driven by learning-by-doing with experience, secular changes over time, and fixed provider-specific effects. We find strong evidence of learning-by-doing: providers become more successful at navigating the cap with experience. This lends larger providers, who gain experience faster, a mechanical advantage.

Taken together, our findings suggest that soft spending limits work, but with important tradeoffs. On the one hand, they do screen out wasteful spending—spending limits allow principals to strike a balance between saving money while still allowing high-value spending to occur. In our setting, this screening stems entirely from regulator scrutiny rather than by encouraging better self-screening by agents. On the other hand, spending limits that require documentation also introduce perhaps undesirable horizontal inequity. We find that employing a documentation requirement gives larger providers a substantial advantage simply because they have more opportunities to learn how to comply with the policy. The result is an efficiency-equity tradeoff: while soft limits allow savings to be targeted, the associated paperwork introduces horizontal inequity by also screening on agents' administrative capacity.

References

Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A. Olken, Ririn Purnamasari, and Matthew Wai-Poi, "Self-Targeting: Evidence from a Field Experiment in Indonesia," *Journal of Political Economy*, April 2016, *124* (2), 371–427. Publisher: The University of Chicago Press.

Amico, Peter, Gregory C. Pope, Poonam Pardasaney, Ben Silver, Jill A. Dever, Ann Meadow, and Pamela West, "Refinements of the Medicare Outpatient Therapy Annual Expenditure Limit Policy," *Physical Therapy*, December 2015, *95* (12), 1638–1649.

APTA, "MedPAC Releases Revised Medicare 'Payment Basics'," Technical Report October 2020.

Brot-Goldberg, Zarek C., Samantha Burn, Timothy Layton, and Boris Vabson, "Rationing Medicine Through Bureaucracy: Authorization Restrictions in Medicare," January 2023.

Burn, Samantha and Ljubica Ristovska, "Informative ordeals in healthcare: Prior authorization of drugs in Medicaid," January 2025.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer, "The Effect of Minimum Wages on Low-Wage Jobs*," *The Quarterly Journal of Economics*, August 2019, *134* (3), 1405–1454.

CFR, "Initial Determinations," 2009.

_ , "Approval of the justification," 2020.

CMS, "Outpatient Therapy Caps: Exceptions Process Required by the DRA," February 2006.

_ , "Therapy Caps Exception Process," Technical Report Transmittal 855, Centers for Medicare and Medicaid February 2006.

_ , "Use of the KX Modifier on Claims Submitted to the Fiscal Intermediary When Some Services Exceed the Therapy Caps | Guidance Portal," HHS-0938-2006-F-8208, Centers for Medicare and Medicaid June 2006.

_ , "Therapy Caps and Advance Beneficiary Notice of Noncoverage (ABN), Form CMS-R-131, FAQs April 2013," Technical Report, Centers for Medicare and Medicaid April 2013.

_ , "An Introduction to: Medicare EHR Incentive Program for Eligible Professionals," Technical Report April 2014.

_ , "Outpatient Therapy Services and Advance Beneficiary Notice of Noncoverage (ABN), Form CMS-R-131, August 2018," August 2018.

_ , "Outpatient Physical and Occupational Therapy Services (L33631)," Technical Report, Centers for Medicare and Medicaid 2020.

_ , "Medicare Benefit Policy Manual. Chapter 15 – Covered Medical and Other Health Services," Technical Report, Centers for Medicare and Medicaid 2024.

_ , "Targeted Probe and Educate | CMS," Technical Report, Centers for Medicare and Medicaid September 2024.

Currie, Janet, "The take-up of social benefits," in "Public Policy and the Distribution of Income," Russell Sage Foundation, 2006, pp. 80–148.

Deshpande, Manasi and Yue Li, "Who Is Screened Out? Application Costs and the Targeting of Disability Programs," *American Economic Journal: Economic Policy*, November 2019, *11* (4), 213–248.

Diamond, Rebecca and Petra Persson, "The Long-term Consequences of Teacher Discretion in Grading of High-stakes Tests," April 2016.

Dillender, Marcus, "What happens when the insurer can say no? Assessing prior authorization as a tool to prevent high-risk prescriptions and to lower costs," *Journal of Public Economics*, September 2018, *165*, 170–200.

DOL, "How Do I File for Unemployment Insurance?," Technical Report, U.S. Department of Labor 2025.

Dunn, Abe, Joshua D Gottlieb, Adam Hale Shapiro, Daniel J Sonnenstuhl, and Pietro Tebaldi, "A Denial a Day Keeps the Doctor Away*," *The Quarterly Journal of Economics*, February 2024, *139* (1), 187–233.

DynCorp, "Study and Report on Outpatient Therapy Utilization: Physical Therapy, Occupational Therapy, and Speech-Language Pathology Services Billed to Medicare Part B in All Settings in 1998, 1999, and 2000," September 2002.

Eliason, Paul, Riley League, Jetson Leder-Luis, Ryan McDevitt, and James Roberts, "Ambulance Taxis: The Impact of Regulation and Litigation on Health Care Fraud," *Journal of Political Economy*, November 2024. Publisher: The University of Chicago Press.

Evans, William N., Shawna Kolka, James X. Sullivan, and Patrick S. Turner, "Fighting Poverty One Family at a Time: Experimental Evidence from an Intervention with Holistic, Individualized, Wrap-Around Services," *American Economic Journal: Economic Policy*, 2024.

FEMA, "Insurance Documentation Required During FEMA Application Process | FEMA.gov," Technical Report, Federal Emergency Management Agency March 2021.

Finkelstein, Amy and Matthew J Notowidigdo, "Take-Up and Targeting: Experimental Evidence from SNAP*," *The Quarterly Journal of Economics*, August 2019, *134* (3), 1505–1556.

Goldstein, Amy, "Medicare Cutbacks Prove Painful to Some," *Washington Post*, May 1999.

Howard, David H. and Ian McCarthy, "Deterrence effects of antifraud and abuse enforcement in health care," *Journal of Health Economics*, January 2021, *75*, 102405.

Hoynes, Hilary W., Nicole Maestas, and Alexander Strand, "Legal Representation in Disability Claims," March 2022.

Ida, Takanori, Takunori Ishihara, Koichiro Ito, Daido Kido, Toru Kitagawa, Shosei Sakaguchi, and Shusaku Sasaki, "Choosing Who Chooses: Selection-Driven Targeting in Energy Rebate Programs," September 2022.

IRS, "SOI Tax Stats - Individual income tax statistics - ZIP Code data (SOI) | Internal Revenue Service," Technical Report, Internal Revenue Service 2025.

Kleven, Henrik Jacobsen, "Bunching," *Annual Review of Economics*, 2016, *8* (1), 435–464. _eprint: https://doi.org/10.1146/annurev-economics-080315-015234.

_ **and Wojciech Kopczuk**, "Transfer Program Complexity and the Take-Up of Social Benefits," *American Economic Journal: Economic Policy*, February 2011, *3* (1), 54–90.

Kopczuk, Wojciech and Cristian Pop-Eleches, "Electronic filing, tax preparers and participation in the Earned Income Tax Credit," *Journal of Public Economics*, August 2007, *91* (7), 1351–1367.

League, Riley, "Administrative Burden and Consolidation in Health Care: Evidence from Medicare Contractor Transitions," 2023.

Leder-Luis, Jetson, "Can Whistleblowers Root Out Public Expenditure Fraud? Evidence from Medicare," *Review of Economics and Statistics*, 2023, *Forthcoming*, 60.

Lieber, Ethan M. J. and Lee M. Lockwood, "Targeting with In-Kind Transfers: Evidence from Medicaid Home Care," *American Economic Review*, April 2019, *109* (4), 1461–1485.

Luthra, Shefali, "Tucked into the budget deal, long-awaited gifts to some health-care providers," *Washington Post*, March 2018.

Macambira, Danil, Michael Geruso, Anthony Lollo, Chima D. Ndumele, and Jacob Wallace, "The Private Provision of Public Services: Evidence from Random Assignment in Medicaid," August 2022.

Mullainathan, Sendhil and Ziad Obermeyer, "Does Machine Learning Automate Moral Hazard and Error?," *American Economic Review*, May 2017, *107* (5), 476–480.

Nichols, Albert L. and Richard J. Zeckhauser, "Targeting Transfers through Restrictions on Recipients," *The American Economic Review*, 1982, *72* (2), 372–377. Publisher: American Economic Association.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, October 2019, *366* (6464), 447–453. Publisher: American Association for the Advancement of Science.

OIG, "A South Texas Physical Therapist Claimed Unallowable Medicare Part B Reimbursement for Outpatient Physical Therapy Services," Technical Report A-06-14-00064, Office of the Inspector General June 2016.

_ , "Fox Rehabilitation Claimed Unallowable Medicare Reimbursement for Outpatient Therapy Services," Technical Report A-02-16-01004, Office of the Inspector General August 2017.

_ , "Many Medicare Claims for Outpatient Physical Therapy Services Did Not Comply With Medicare Requirements," Technical Report A-05-14-00041, Office of the Inspector General March 2018.

O'Malley, A. James, Thomas A. Bubolz, and Jonathan S. Skinner, "The diffusion of health care fraud: A bipartite network analysis," *Social Science & Medicine (1982)*, June 2023, *327*, 115927.

Rafkin, Charlie, Adam Solomon, and Evan Soltas, "Self-Targeting in U.S. Transfer Programs," September 2023.

Shepard, Mark and Myles Wagner, "Do Ordeals Work for Selection Markets? Evidence from Health Insurance Auto-Enrollment," August 2024.

Shi, Maggie, "Monitoring for Waste: Evidence from Medicare Audits*," *The Quarterly Journal of Economics*, May 2024, *139* (2), 993–1049.

SSA, "Disability Evaluation Under Social Security," Technical Report 64-039, Social Security Administration 2008.

US GAO, "Early Resolution of Overcharges for Therapy in Nursing Homes Is Unlikely," Technical Report GAO/HEHS-96-145 August 1996.

WebPT, "CMS Final Rule: The Countdown to 2024 Begins," 2024.

Woodward, Susan E. and Robert E. Hall, "Diagnosing Consumer Confusion and Suboptimal Shopping Effort: Theory and Mortgage-Market Evidence," *American Economic Review*, December 2012, *102* (7), 3249–3276.

Zeckhauser, Richard, "Strategic sorting: the role of ordeals in health care," *Economics & Philosophy*, March 2021, *37* (1), 64–81.

Zwick, Eric, "The Costs of Corporate Tax Complexity," *American Economic Journal: Economic Policy*, May 2021, *13* (2), 467–500.

A Appendix Tables

Table A1: Highest-Volume In-Office Physical Therapy Procedures

HCPCS	Description	N. Lines (1000s)	Share of Beneficiaries (%)
97110	Therapeutic exercises	3065.8	20.3
97140	Manual therapy	1657.7	12.8
G0283	Elec stimulation other than wound	875.5	7.4
97035	Ultrasound therapy	843.5	8.3
97112	Neuromuscular reeducation	651.9	5.4
97530	Therapeutic activities	640.7	5.5
97032	Electrical stimulation	415.5	3.5
97001	PT evaluation	279.8	18.6
97124	Massage therapy	226.8	2.0
97116	Gait training therapy	184.2	1.8
97010	Hot or cold packs therapy	130.9	1.3
97012	Mechanical traction therapy	102.1	1.1
97150	Group therapeutic procedures	85.2	1.0
97113	Aquatic therapy/exercises	84.8	0.7
97535	Self care management training	76.2	1.5

This table reports the 15 highest volume PT procedures in the 2006 20% Carrier sample. "N. Lines" is the number of lines billed for each procedure code (HCPCS). "Share of beneficiaries" is the share of beneficiaries in the bunching sample that ever receive that procedure in that year. Data: 20% Medicare Carrier claims.

Table A2: Robustness: Regression Results on Screening and Differences by Provider Size using Age, 2005-2006

	(1)	(2)	(3)	(4)	(5)	(6)
	Denied | Attempt		Deterred | Approach		Stop at Cap | Approach	
Age	.668	.664	-.796	-.755	.0271	.0534
	(.613)	(.613)	(2.37)	(2.36)	(2.72)	(2.72)
Age × 2006	-27.5***	-27.3***	1.06	1.02	-18.6***	-18.7***
	(4.54)	(4.31)	(3.84)	(3.85)	(5.06)	(4.91)
Small provider		.089		-1.18**		-.752
		(.145)		(.55)		(.63)
Small provider × 2006		12.4***		-.905		9.87***
		(1.05)		(.874)		(1.22)
Outcome mean, 2005	1.0	1.0	18.6	18.6	21.2	21.2
Outcome mean, 2006	22.1	22.1	26.5	26.5	51.9	51.9
Week-year FE	X	X	X	X	X	X
Cluster	Provider	Provider	Provider	Provider	Provider	Provider
N. Providers	10623	10623	11535	11535	11535	11535
N. Beneficiaries	47919	47919	63238	63238	63238	63238
N. Observations	103598	103598	72002	72002	72002	72002

* $p < .10$, ** $p < .05$,*** $p < .01$. This table presents the coefficients from estimating equation (3) with log patient age as X_i (columns (1), (3), and (5)) and equation (4) with log patient age as X_i and an indicator for whether a patient goes to a small provider as C_i (columns (2), (4), and (6)). Outcomes are all defined at the patient-level. The regression is clustered at the provider (TIN-state) level. All specifications restrict to patients over 65. Columns (1)-(2) restrict to the sample of patients who ever make an attempt to go past the cap and columns (3)-(6) restrict to the sample of patients who ever approach the cap, as defined in Section 2.3. Data: 20% Medicare Carrier claims.

Table A3: Regression Results on Screening and Differences by Patient Demographics, 2005-2006

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Denied | Attempt			Deterred | Approach			Stop at Cap | Approach		
Predicted Spending	-1.6***	-1.62***	-1.61***	-17***	-17.1***	-17.1***	-20.7***	-20.9***	-20.9***
	(.188)	(.189)	(.194)	(.706)	(.707)	(.722)	(.803)	(.812)	(.834)
Predicted Spending × 2006	-7.08***	-6.98***	-6.98***	-.297	-.0484	.00384	-3.76**	-4.01***	-3.96***
	(1.34)	(1.37)	(1.37)	(1.04)	(1.04)	(1.05)	(1.49)	(1.49)	(1.5)
Non-white	.54**			.169			.892		
	(.275)			(.614)			(.747)		
Non-white × 2006	8.88***			-2.72***			5.8***		
	(1.46)			(.955)			(1.36)		
Dual		.437**			.979*			1.5**	
		(.219)			(.578)			(.701)	
Dual × 2006		4.97***			-2.78***			4.37***	
		(1.31)			(.929)			(1.31)	
Low-income subsidy			.401*			.323			.749
			(.21)			(.569)			(.671)
Low-income subsidy × 2006			4.51***			-2.45***			4.5***
			(1.26)			(.913)			(1.28)
Outcome mean, 2005	1.0	1.0	1.0	18.4	18.4	18.3	21.1	21.1	21.0
Outcome mean, 2006	23.1	23.1	23.1	26.3	26.3	26.3	52.2	52.2	52.2
Week-year FE	X	X	X	X	X	X	X	X	X
Cluster	Provider	Provider	Provider	Provider	Provider	Provider	Provider	Provider	Provider
N. Providers	11064	11064	11041	11911	11911	11887	11911	11911	11887
N. Beneficiaries	53560	53560	52986	70518	70518	69788	70518	70518	69788
N. Observations	116360	116360	115228	80532	80532	79782	80532	80532	79782

* $p < .10$, ** $p < .05$,*** $p < .01$. This table presents the coefficients from estimating equation (4) with log predicted PT spending as X_i and an indicator for a patient's demographic characteristic C_i: non-white (columns (1), (4), and (7)), dual-eligibility for Medicaid (columns (2), (5), and (8)), and Part D Low Income Subsidy status in 2006 (columns (3), (6), and (9)). Outcomes are all defined at the patient-level. All specifications include week-year fixed effects for the calendar week(s) that a beneficiary attempts or approaches the cap. The regression is clustered at the provider (TIN-state) level. Columns (1)-(3) restrict to the sample of 2005-2006 patients who ever make an attempt to go past the cap, while columns (4)-(9) restrict to patients who ever approach the cap, as defined in Section 2.3. Data: 20% Medicare Carrier claims.

Table A4: Regression Results on Screening and Differences by Patient Demographics with Provider-Year FEs, 2005-2006

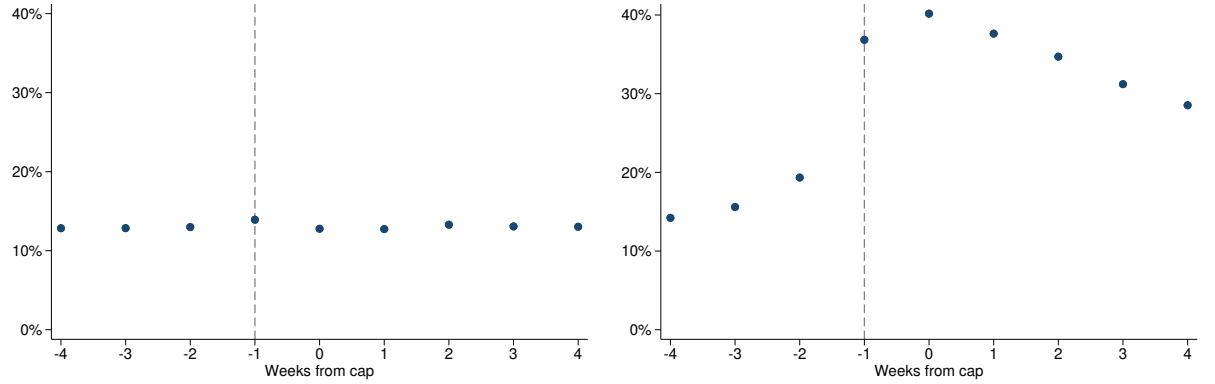
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Denied | Attempt				Deterred | Approach				Stop at Cap | Approach		
Predicted Spending	-.528***	-.528***	-.532***	-.517***	-13.5***	-13.5***	-13.6***	-13.6***	-14.5***	-14.5***	-14.6***	-14.6***
	(.147)	(.147)	(.147)	(.148)	(.856)	(.856)	(.855)	(.873)	(.894)	(.894)	(.893)	(.917)
Predicted Spending × 2006	-4.84***	-4.87***	-4.85***	-4.88***	.492	.488	.521	.545	-2.64*	-2.64*	-2.62*	-2.61*
	(.843)	(.842)	(.842)	(.841)	(1.26)	(1.26)	(1.26)	(1.27)	(1.4)	(1.4)	(1.4)	(1.42)
Non-white		-.212				.874				.6		
		(.189)				(.806)				(.853)		
Non-white × 2006		1.99**				-1.62				-.431		
		(.852)				(1.2)				(1.26)		
Dual			.0943				.867				1.15	
			(.21)				(.751)				(.836)	
Dual × 2006			.358				-.875				-.33	
			(.769)				(1.15)				(1.29)	
Low-income subsidy				.0226				.168				.358
				(.199)				(.741)				(.813)
Low-income subsidy × 2006				.665				-.944				.43
				(.779)				(1.14)				(1.27)
Outcome mean, 2005	0.9	0.9	0.9	0.9	18.0	18.0	18.0	18.0	20.9	20.9	20.9	20.9
Outcome mean, 2006	22.4	22.4	22.4	22.4	24.9	24.9	24.9	24.9	52.2	52.2	52.2	52.2
Week-year FE	X	X	X	X	X	X	X	X	X	X	X	X
Provider-year FE	X	X	X	X	X	X	X	X	X	X	X	X
Cluster	Provider	Provider	Provider	Provider	Provider	Provider	Provider	Provider	Provider	Provider	Provider	Provider
N. Providers	10788	10788	10788	10764	9725	9725	9725	9700	9725	9725	9725	9700
N. Beneficiaries	52968	52968	52968	52394	65853	65853	65853	65130	65853	65853	65853	65130
N. Observations	115716	115716	115716	114584	75528	75528	75528	74783	75528	75528	75528	74783

* $p < .10$, ** $p < .05$,*** $p < .01$. This table presents the coefficients from estimating equation (3) with log predicted PT spending as X_i (columns (1) and (5)) and the inclusion of provider-year fixed effects, and equation (4) with log predicted PT spending as X_i, the inclusion of provider-year fixed effects, and an indicator for a patient's demographic characteristic C_i: non-white (columns (2), (6), and (10)), dual-eligibility for Medicaid (columns (3), (7) and (11)), and Part D Low Income Subsidy status in 2006 (columns (4), (8), and (12)). Outcomes are all defined at the patient-level. All specifications include week-year fixed effects for the calendar week(s) that a beneficiary attempts or approaches the cap. The regression is clustered at the provider (TIN-state) level. Columns (1)-(4) restrict to the sample of 2005-2006 patients who ever make an attempt to go past the cap, while columns (5)-(12) restrict to patients who ever approach the cap, as defined in Section 2.3. Data: 20% Medicare Carrier claims.

B Appendix Figures

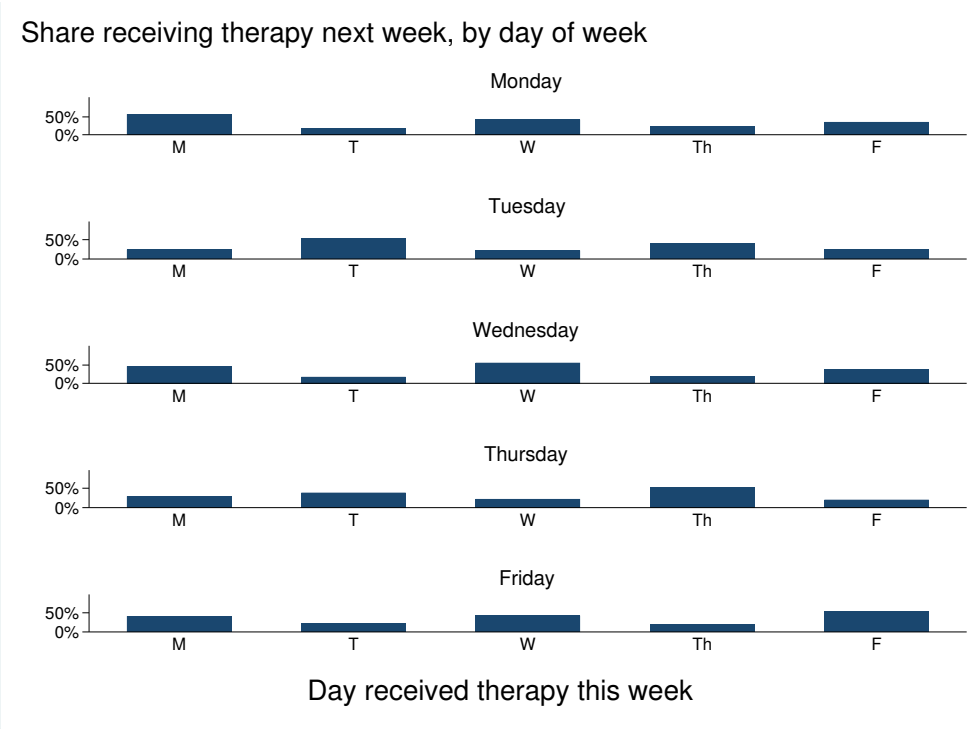
Figure B1: Claim-level Denial Rates by Weeks to Cap, Pre- and Post-Reform

(a) Pre-Reform, 2005 **(b) Post-Reform, 2006**



This figure plots the average denial rate for PT claims depending on how far the patient is from the therapy cap, where the denial rate is defined at the claim-level and is the share of claims with at least one denied line. Panel (a) graphs the denial rates relative to a (placebo) 2006 cap in 2005, and panel (b) graphs the denial rates relative to the cap in 2006. A patient's distance to the cap is calculated in terms weeks of care. When "weeks from cap" is negative, it denotes how many additional weeks of care a patient would have to receive to reach the cap. When "weeks from cap" is positive, it denotes how many weeks ago the patient passed the cap. The procedure to define "weeks from cap" is described in detail in Section 2.3. Data: 20% Medicare Carrier claims.

Figure B2: Share Receiving Therapy on Each Day of Week

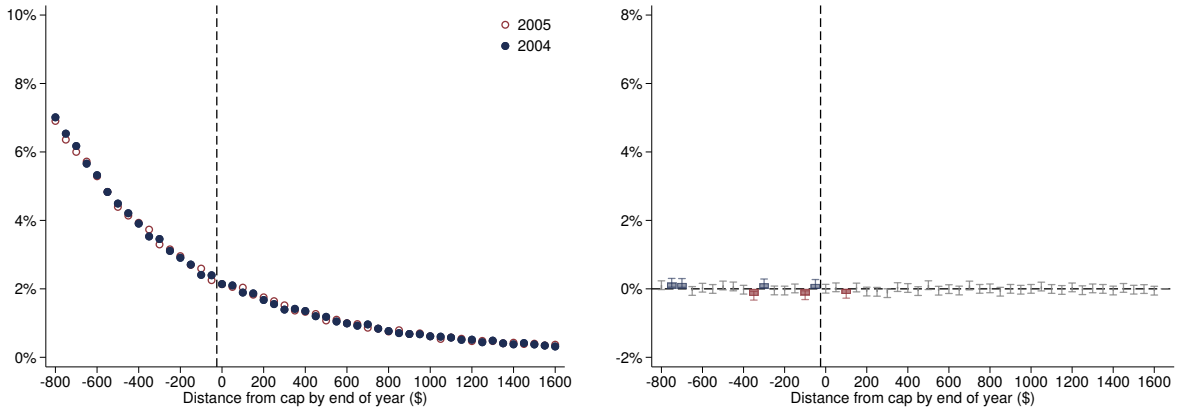


This plot graphs the share of PT visits on each day of the week, given the day of therapy in the previous week. Data: 20% Medicare Carrier claims.

Figure B3: Distributions of Spending Around Cap in 2004 and 2005

(a) Distribution of Spending

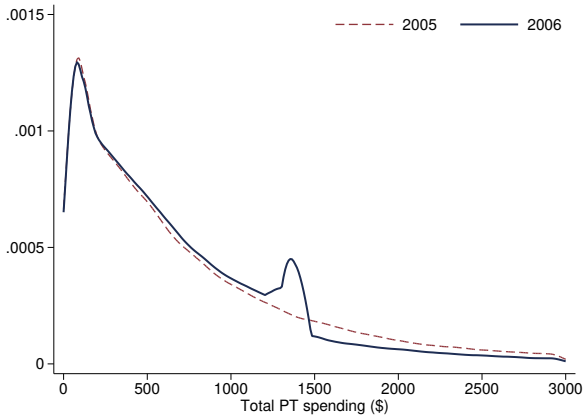
(b) Difference between 2005 and 2004



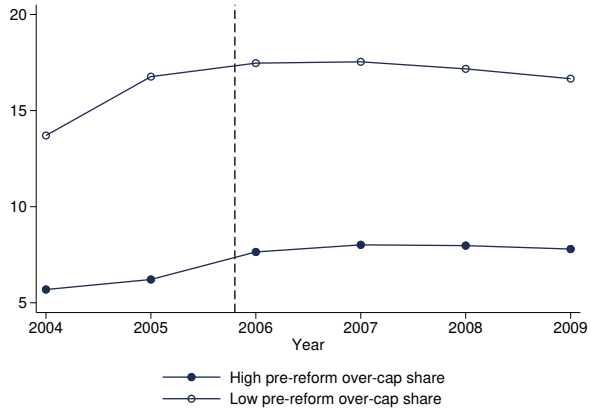
This figure plots the (a) distributions of end-of-year physical therapy spending around the cap in 2004 and 2005 and (b) the difference in the distributions between 2004 to 2005. Distance from cap is calculated in bins of $50 relative to the 2006 cap and shares are calculated as the share of patients within [-$800, $1600] of the cap. Data: 20% Medicare Carrier claims.

Figure B4: Extensive Margin Responses

(a) Distribution of PT Spending, 2005-2006



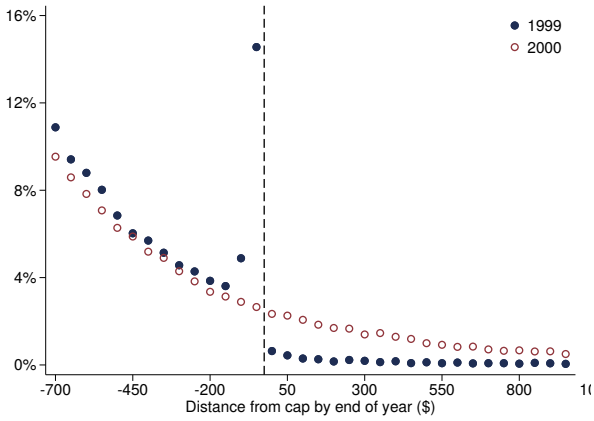
(b) Mean beneficiaries per provider, by 2005 over cap share



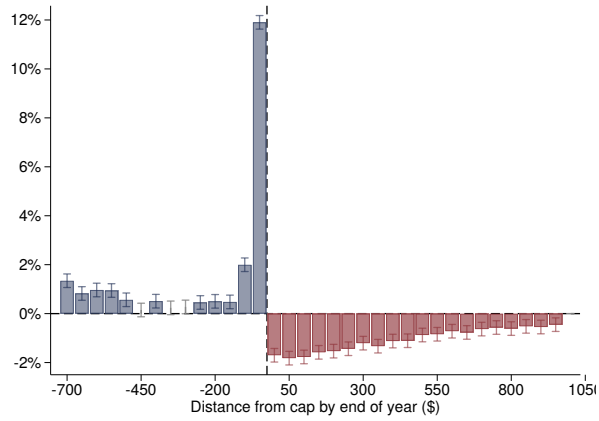
This figure characterizes potential extensive margin responses to the 2006 therapy cap. Panel (a) plots a kernel density of Medicare physical therapy spending in 2005 and 2006 up to $3000. Since Medicare pays for 80 percent of allowed charges and patients are responsible for the remaining 20 percent, so the cap appears at $0.8 \times \$1740 = \$1392$. Panel (b) plots the average number of beneficiaries per provider, split by whether a provider had a high or low over (placebo) cap share in 2005. Data: 20% Medicare Carrier claims.

Figure B5: Bunching in Dollars, 1999-2000

(a) 1999 and 2000 Spending Distributions

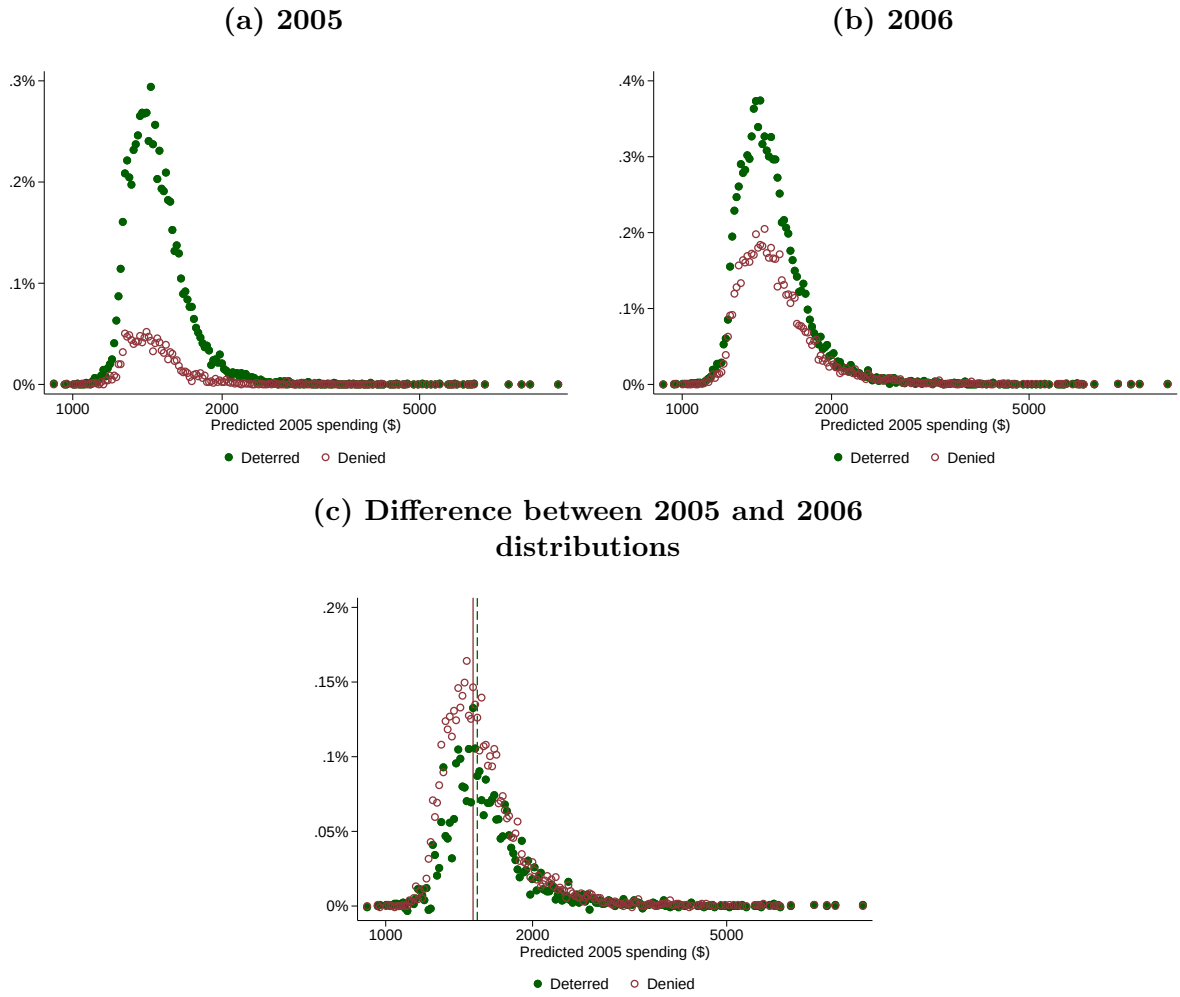


(b) Difference between 1999 and 2000



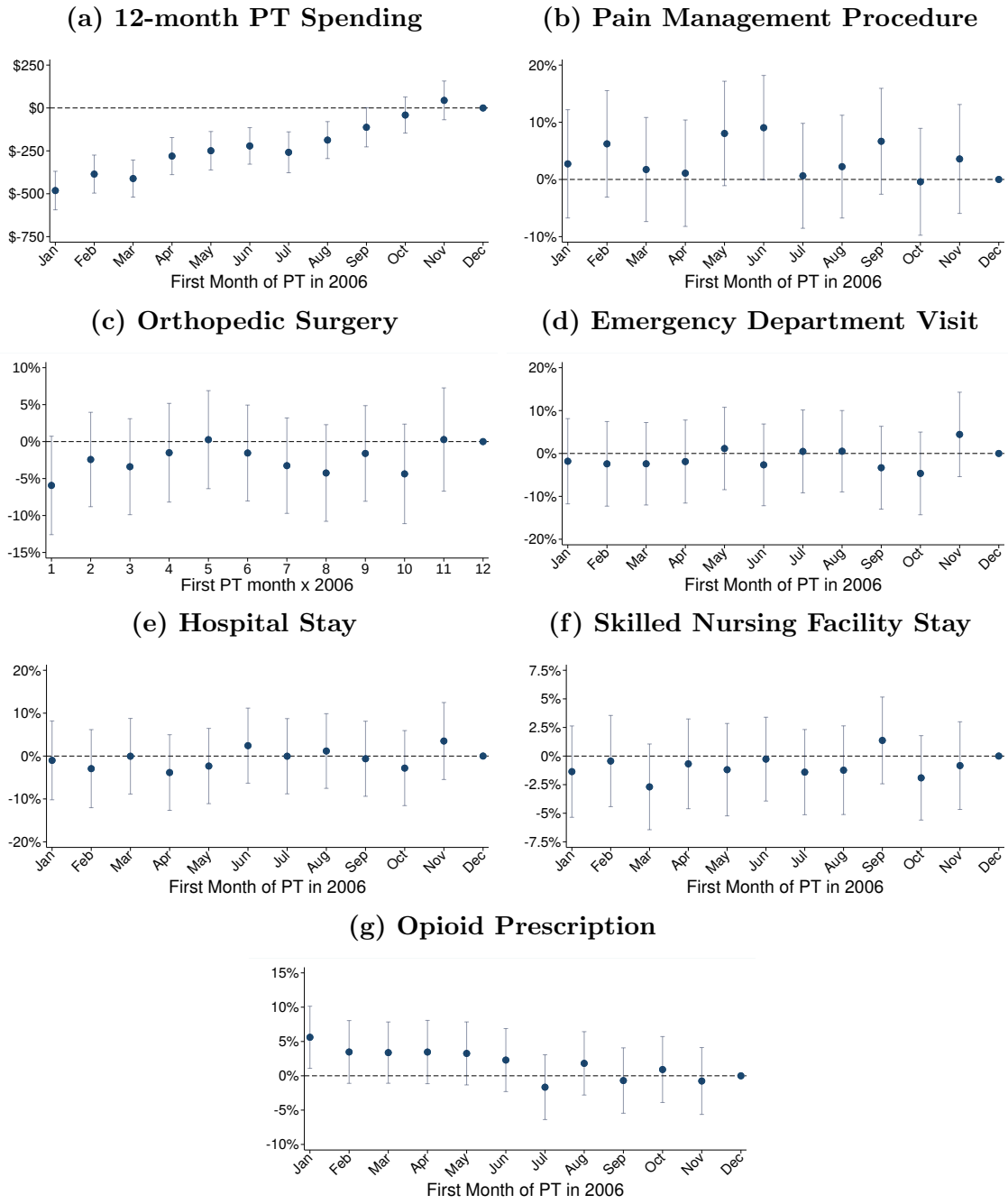
This figure plots the (a) distributions of end-of-year physical therapy spending around the cap in 1999 and 2000 and (b) the difference in the distributions between 1999 and 2000. Distance from cap is calculated in bins of $50 relative to the 1999 cap (inflation-adjusted to 2005 dollars) and shares are calculated as the share of patients within [-$700, $1400] of the cap. Data: 20% Medicare Carrier claims.

Figure B6: Distribution of predicted PT spending, by deterred and denied



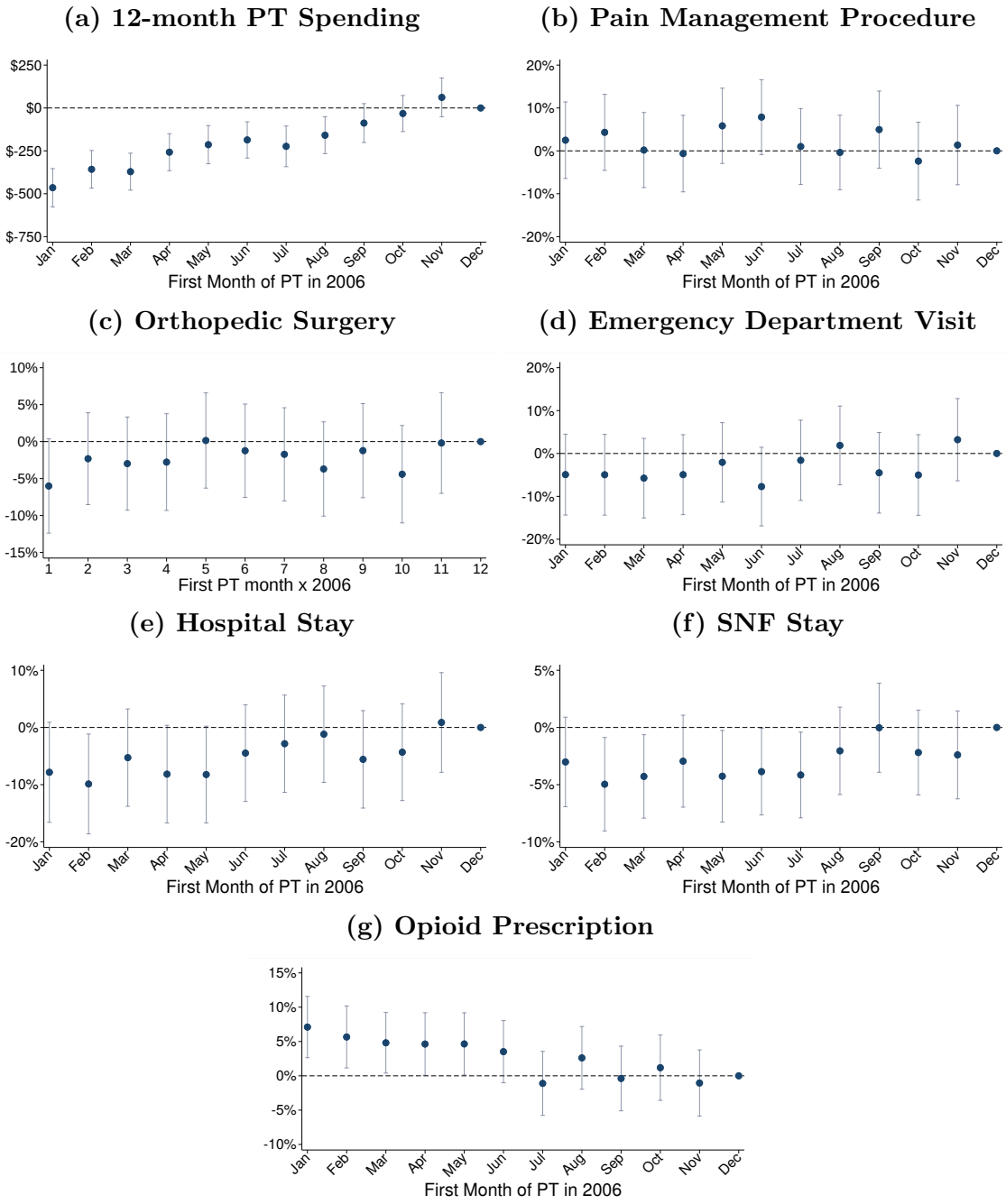
This figure plots the distribution of predicted 2005 PT spending for "deterred" patients, and "denied" patients. Panel (a) plots the distribution of predicted spending for each set of patients in 2005, panel (b) plots the distributions in 2006, and panel (c) plots the difference between the 2006 and 2005 distributions, as well as the medians of each distribution (solid line for denied and dotted line for deterred). Section 2.3 describes the construction of the weeks from cap measure as well as the definitions of deterrence and denial in further detail, and Section 4 and Appendix Section D describe the predicted PT spending measure in further detail. Data: 20% Medicare Carrier claims.

Figure B7: Reduced Form Spending and Health Outcomes, Dual Eligibles



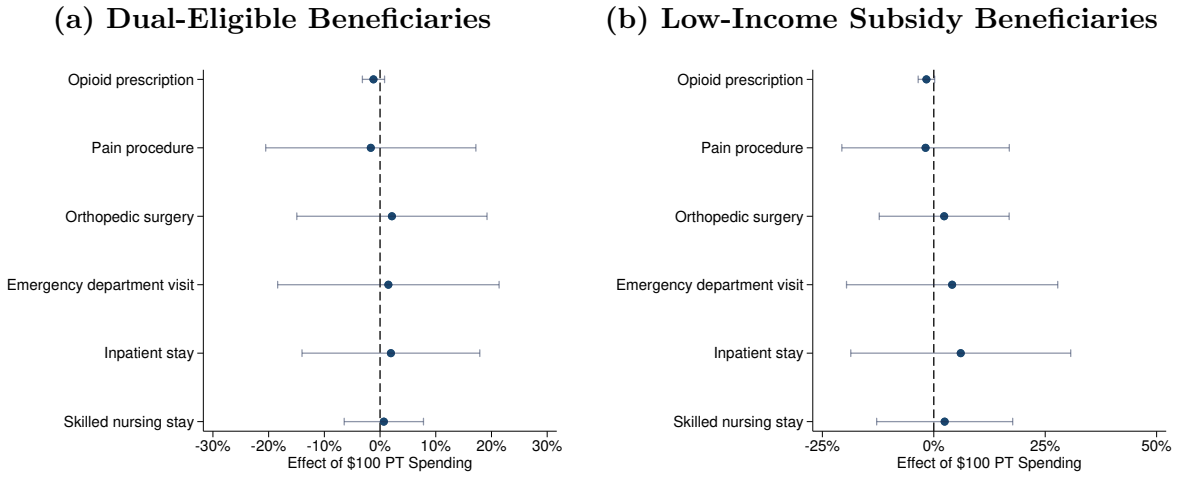
This figure plots the coefficient θ_f, which denotes the interaction between an indicator for 2006 and an indicator for month of first PT, from equation (1). Sample is restricted to beneficiaries with more than $200 in 12-month PT spending who are dually-eligible for Medicare and Medicaid. Panel (a) plots the coefficients on 12-month PT spending ($). Panel (b) plots the coefficients on an indicator for pain management procedures, panel (c) plots the coefficients on an indicator for orthopedic surgery, panel (d) plots the coefficients on an indicator for emergency department visit, panel (e) plots the coefficients on an indicator for a hospital stay, panel (f) plots the coefficients on an indicator for a skilled nursing facility stay, and panel (g) plots the coefficients on an indicator for opioid prescriptions. All outcomes other than opioid prescriptions are measured within 12 months of the first PT session; opioid prescriptions are measured 12-24 months after the first PT session. Section E describes how the health outcome measures are defined. Data: 20% Medicare Carrier, Outpatient, MEDPAR, and Part D claims.

Figure B8: Reduced Form Spending and Health Outcomes, Low Income Subsidy



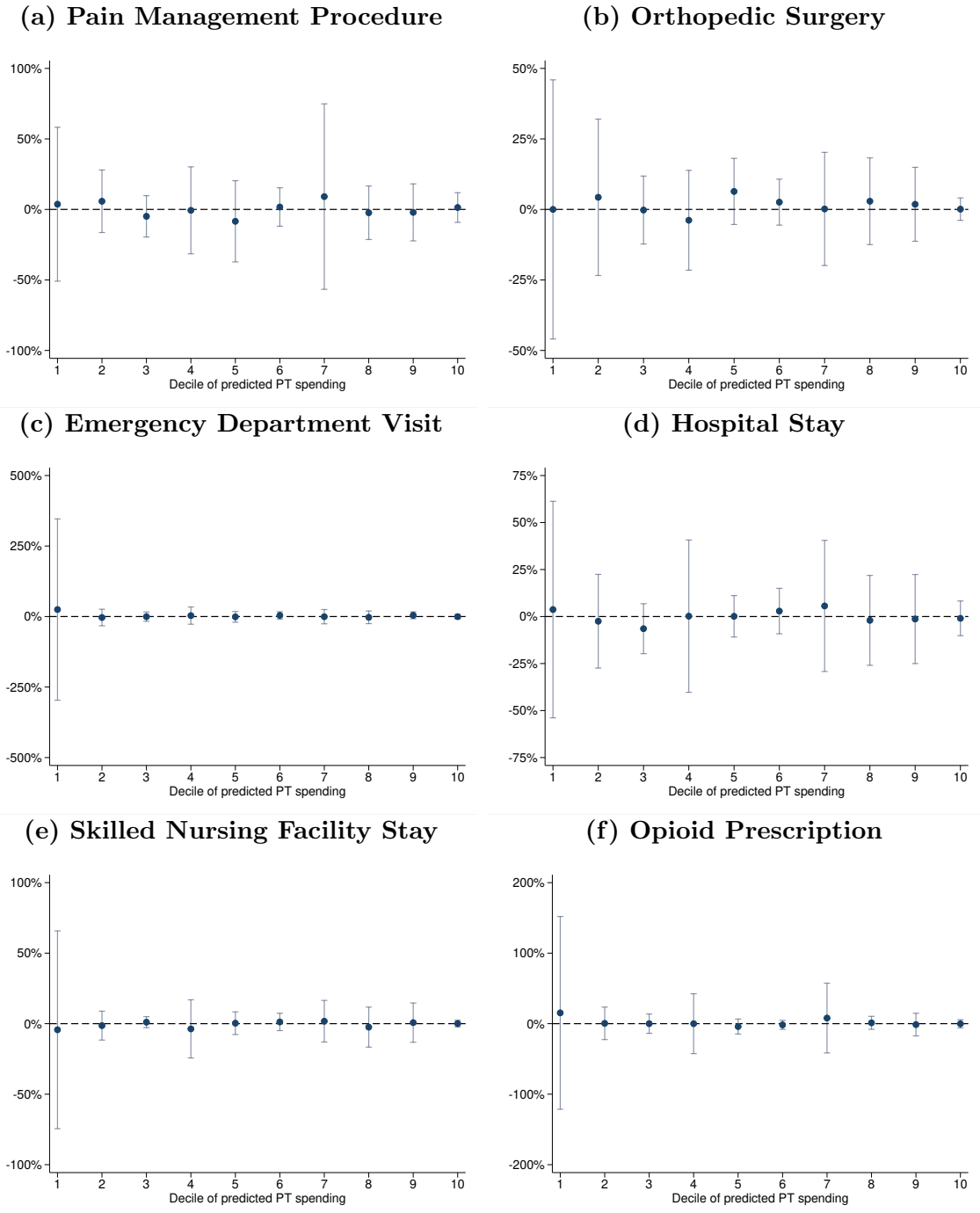
This figure plots the coefficient θ_f, which denotes the interaction between an indicator for 2006 and an indicator for month of first PT, from equation (1). Sample is restricted to beneficiaries with more than $200 in 12-month PT spending who receive the Low Income Subsidy in 2006. Panel (a) plots the coefficients on 12-month PT spending ($). Panel (b) plots the coefficients on an indicator for pain management procedures, panel (c) plots the coefficients on an indicator for orthopedic surgery, panel (d) plots the coefficients on an indicator for emergency department visit, panel (e) plots the coefficients on an indicator for a hospital stay, panel (f) plots the coefficients on an indicator for a skilled nursing facility stay, and panel (g) plots the coefficients on an indicator for opioid prescriptions. All outcomes other than opioid prescriptions are measured within 12 months of the first PT session; opioid prescriptions are measured 12-24 months after the first PT session. Section E describes how the health outcome measures are defined. Data: 20% Medicare Carrier, Outpatient, MEDPAR, and Part D claims.

Figure B9: IV: Effect of PT Spending on Health Outcomes, Low Income Populations



This figure plots the coefficient β, which denotes the effects of an additional \$100 of PT on an indicator for each outcome, from equation (2). Panel (a) subsets to dual-eligible beneficiaries and panel (b) subsets to beneficiaries receiving the Part D Low Income Subsidy (LIS) in 2006. All outcomes other than opioid prescriptions are measured within 12 months of the first PT session; opioid prescriptions are measured 12-24 months after the first PT session. Section E describes how the health outcome measures are defined. Sample is restricted to beneficiaries with more than \$200 in 12-month PT spending. Reduced form results are reported in Figures B7 and B8 20% Medicare Carrier, Outpatient, MEDPAR, and Part D claims.

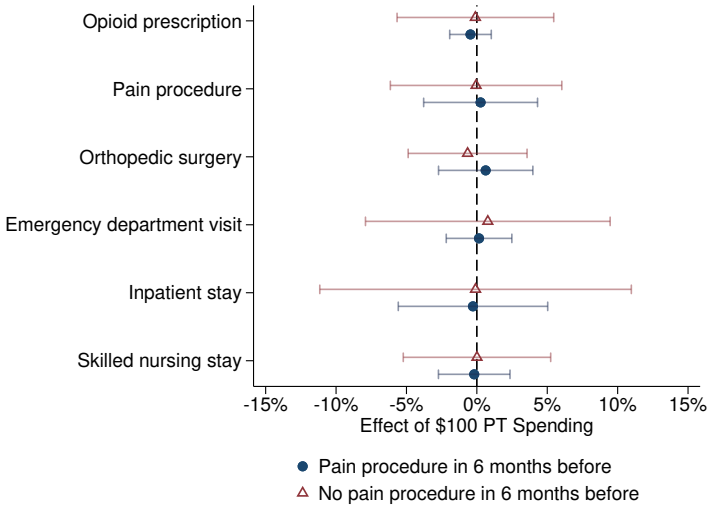
Figure B10: IV: Effect of PT Spending on Health Outcomes, by Decile of Predicted Spending



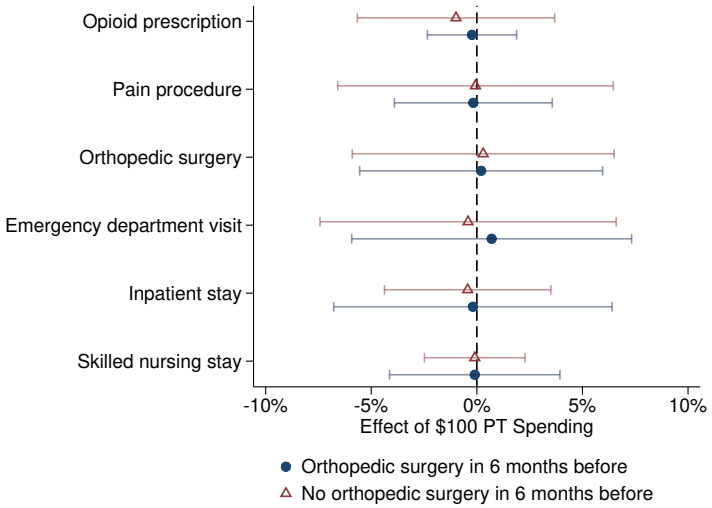
This figure plots the coefficient β, which denotes the effects of an additional \$100 of PT on an indicator of each outcome, from estimating equation (2) on different subsets of predicted need deciles. All outcomes other than opioid prescriptions are measured within 12 months of the first PT session; opioid prescriptions are measured 12-24 months after the first PT session. Section D describes how predicted PT spending is constructed, and Section E describes how the health outcome measures are defined. Sample is restricted to beneficiaries with more than \$200 in 12-month PT spending. Data: 20% Medicare Carrier, Outpatient, MEDPAR, and Part D claims.

Figure B11: IV: Effect of PT Spending on Health Outcomes, By Recent Utilization

(a) By Recent Pain Procedure

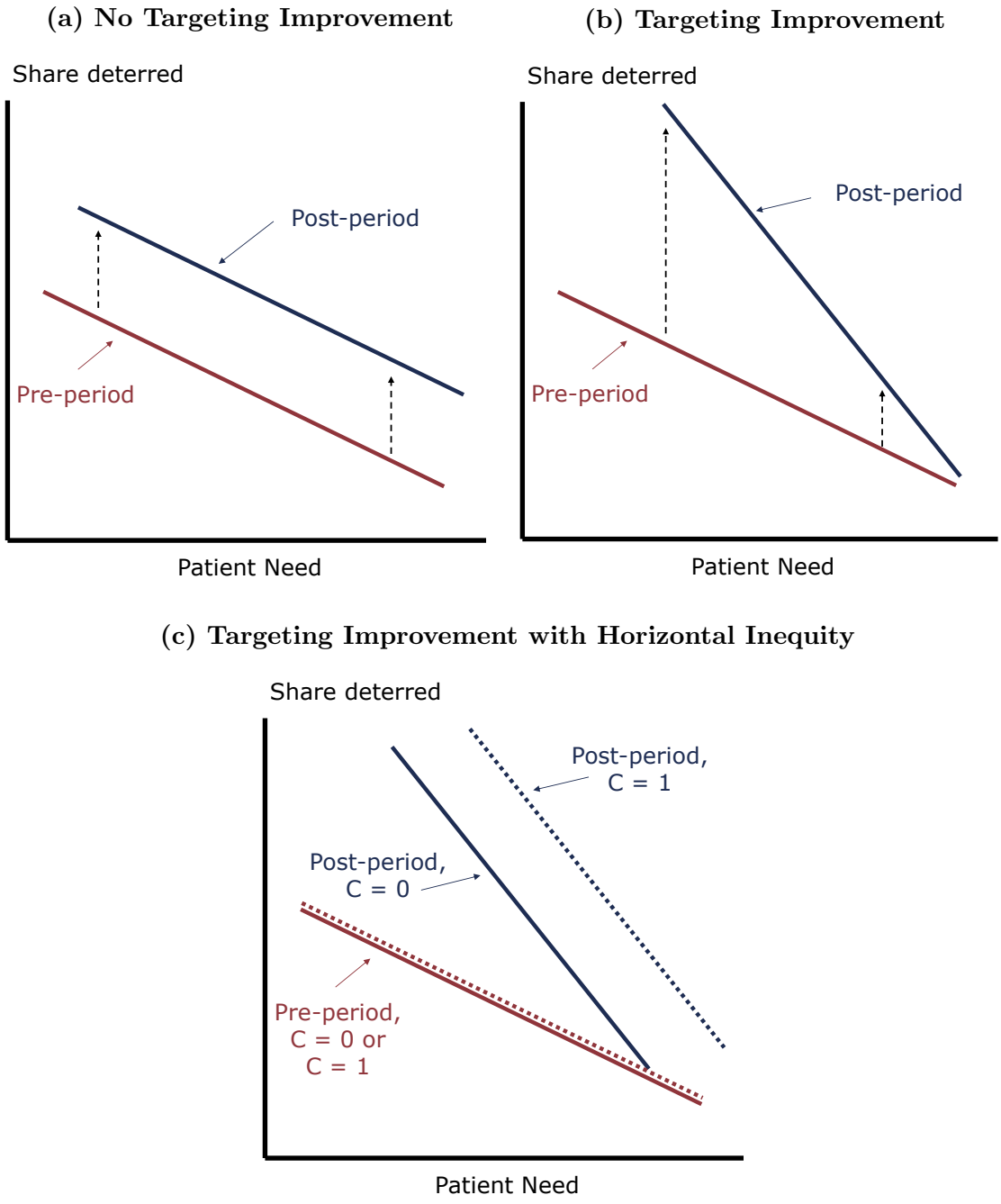


(b) By Recent Orthopedic Surgery



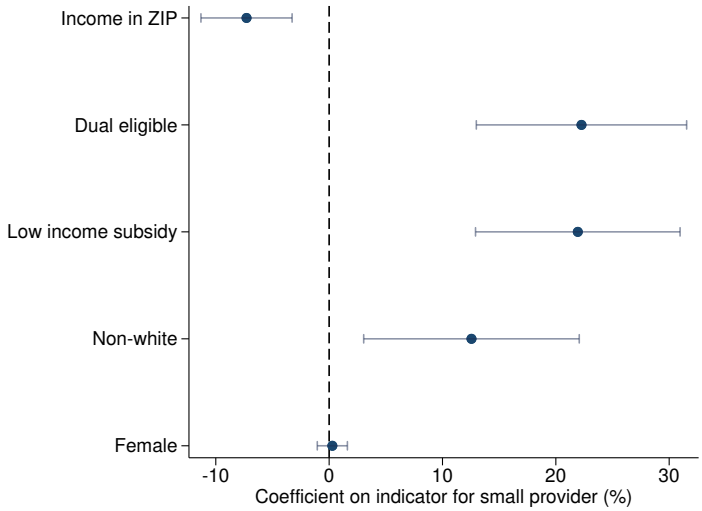
This figure plots the coefficient β, which denotes the effects of an additional $100 of PT on an indicator for each outcome, from equation (2). Results are stratified by whether (a) a patient had a pain procedure in the last 6 months or (b) a patient had an orthopedic surgery in the last 6 months. All outcomes other than opioid prescriptions are measured within 12 months of the first PT session; opioid prescriptions are measured 12-24 months after the first PT session. Section E describes how the health outcome measures are defined. Sample is restricted to beneficiaries with more than $200 in 12-month PT spending. Reduced form results are reported in Figures B7 and B8 20% Medicare Carrier, Outpatient, MEDPAR, and Part D claims.

Figure B12: Illustration of Targeting and Horizontal Inequity Test



This figure presents illustrations of the screening exercise in Figures 6 and 7, using the deterrence rate as an example. Panel (a) illustrates the case when the introduction of the cap increases the share of patients who are deterred, but this increase in deterrence is not targeted. The increase in the share deterred is uniform across all levels of patient need. Panel (b) illustrates the case when the cap increases deterrence, and the increase in deterrence is targeted. In this case, the increase in the share deterred is larger for low-need patients and smaller for high-need patients. Panel (c) illustrates the case when the cap is targeted on need, and also introduces horizontal inequity on characteristic C. Holding fixed patient need, the increase in share stopping at the cap is larger when $C = 1$ than when $C = 0$.

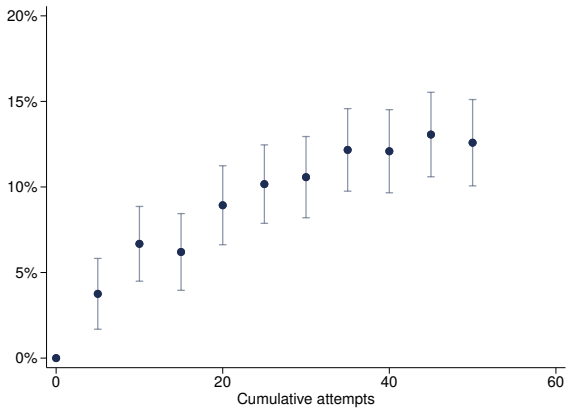
Figure B13: Correlation Between Provider Size and Patient Demographics



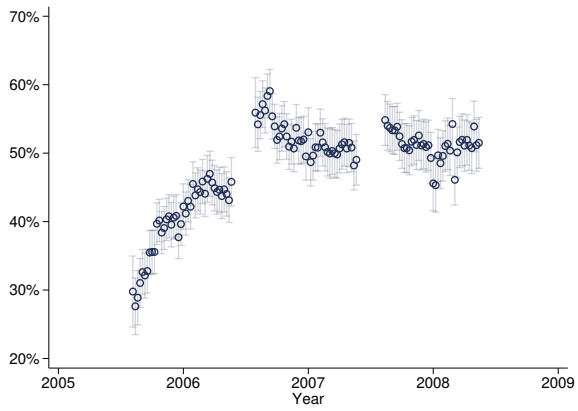
This figure plots the coefficient of the relationship between patient characteristics and an indicator variable for going to a small (below-median size) provider, among patients who approach the cap in 2006 (as defined in Section 2.3). Provider size is measured as a TIN-state's 2006-2008 Medicare patient count. Each regression is clustered at the beneficiary-level. Data: 20% Medicare Carrier claims, Master Beneficiary files, and 2006 Individual Income Tax ZIP Code data (SOI Tax Stats, Internal Revenue Service).

Figure B14: Decomposition of Size Advantage on Approvals

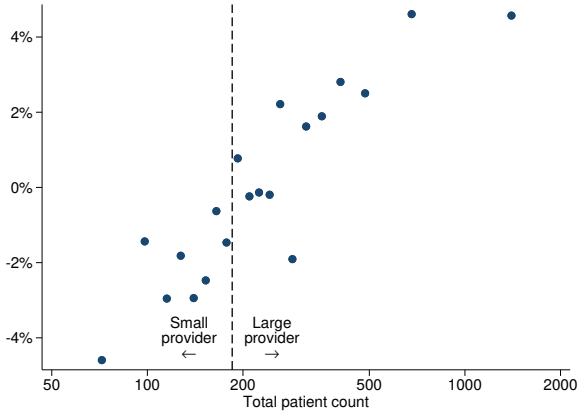
(a) Learning-by-Doing (κ^e)



(b) Industry-Wide Time Trend ($Week_t$)

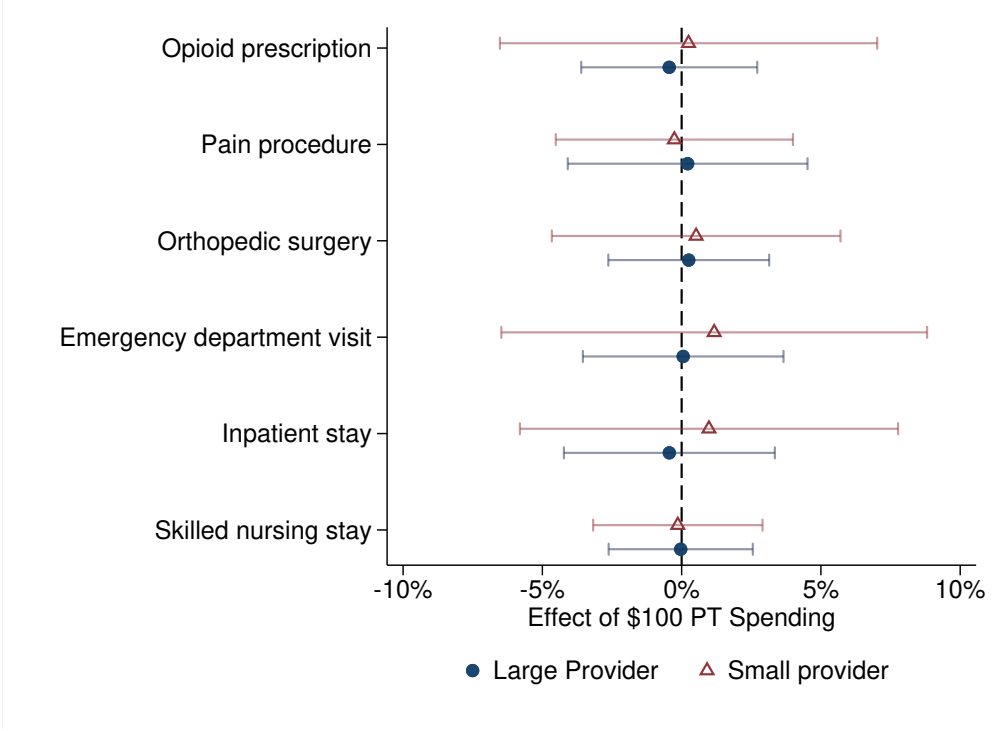


(c) Provider Fixed Effects (α_j) vs. Size



This figure characterizes differences in approval rates on cap attempts by provider size and experience in 2006-2008 using data and regression estimates from equation (5). Panel (a) plots the coefficient on the provider's cumulative number of prior attempts on approvals. Panel (b) provides estimates of weekly industry-wide trends. Panel (c) plots the provider fixed effects against provider patient count. Data: 20% Medicare Carrier claims.

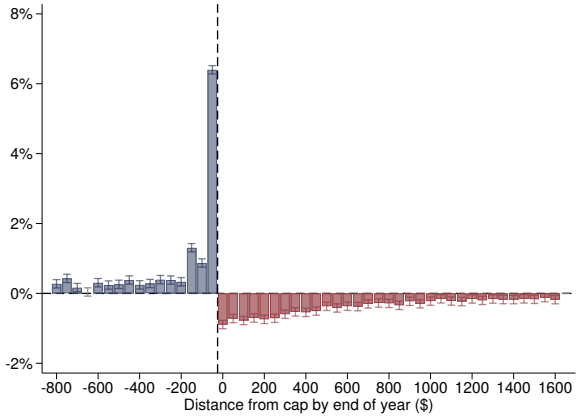
Figure B15: IV: Effect of PT Spending on Health Outcomes, By Provider Size



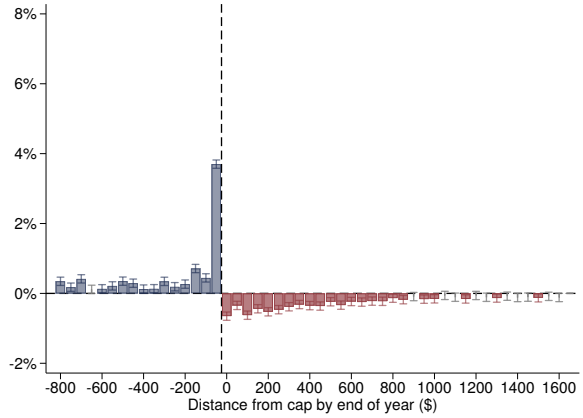
This figure plots the coefficient β, which denotes the effects of an additional \$100 of PT on an indicator for each outcome, from equation (2). Results are stratified by whether a patient goes to an above-median-sized provider, defined as provider who sees 925 or more Medicare patients in 2006-2008. All outcomes other than opioid prescriptions are measured within 12 months of the first PT session; opioid prescriptions are measured 12-24 months after the first PT session. Section E describes how the health outcome measures are defined. Sample is restricted to beneficiaries with more than \$200 in 12-month PT spending. Reduced form results are reported in Figures B7 and B8 20% Medicare Carrier, Outpatient, MEDPAR, and Part D claims.

Figure B16: Distributions of Spending Around Cap in 2006-2008, Relative to 2005

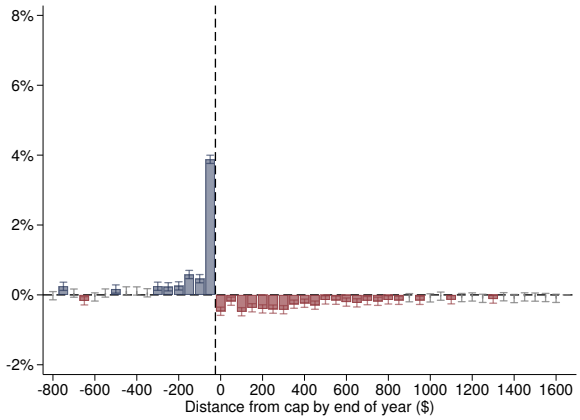
(a) Excess and Missing Mass in 2006



(b) Excess and Missing Mass in 2007



(c) Excess and Missing Mass in 2008



This figure plots the difference in the distributions of PT spending around the cap between 2005 to (a) 2006 (reproduced from Figure 2), (b) 2007, and (c) 2008. Distance from cap is calculated in bins of $50 relative to the 2006 cap and shares are calculated as the share of patients within [-$800, $1600] of the cap. Data: 20% Medicare Carrier claims.

C Policy Context: 1999 Hard Cap

The first therapy cap regime spanned January-December 1999 and was the result of the Balanced Budget Act of 1997. This legislation introduced two separate $1500 caps—one for PT/SLP and one for OT.[27] The 1999 cap was referred to as a "hard cap" in that there was no exceptions process, and Medicare would not cover any services above the cap. Implementation of the cap was imperfect,[28] but the cap was highly salient to providers. As a result of aggressive lobbying by the PT industry,[29] Congress placed a 2-year moratorium on the cap in 2000, and in subsequent years continued to include provisions to extend the delay of the cap one year at a time (until 2006).

Figure B5 plots the 1999 and 2000 spending distributions and differences in distributions in the range from $700 (in 1999 dollars) below the cap to $1300 above the cap, which is equal to approximately $800 and $1600 in 2006 dollars. As depicted in the distributions, some patients do manage to get care reimbursed above the cap in 1999, most likely due to imperfect enforcement of the hard cap. According to a 2003 report on the implementation of the hard cap, CMS could not accurately track cumulative per-patient spending because of "Y2K"-related computing constraints and low provider awareness of the modifier codes introduced to track therapy claims (DynCorp, 2002).

[27] The three services were intended to each have their own cap, but the combination of PT and SLP into one cap is purportedly the result of a missing "oxford comma" in the text of the legislation (WebPT, 2024)

[28] It was revealed in later reports about the 1999 cap that CMS could not accurately track cumulative per-patient spending because of Y2K-related computing constraints and low provider usage of the modifier codes introduced to track therapy claims.

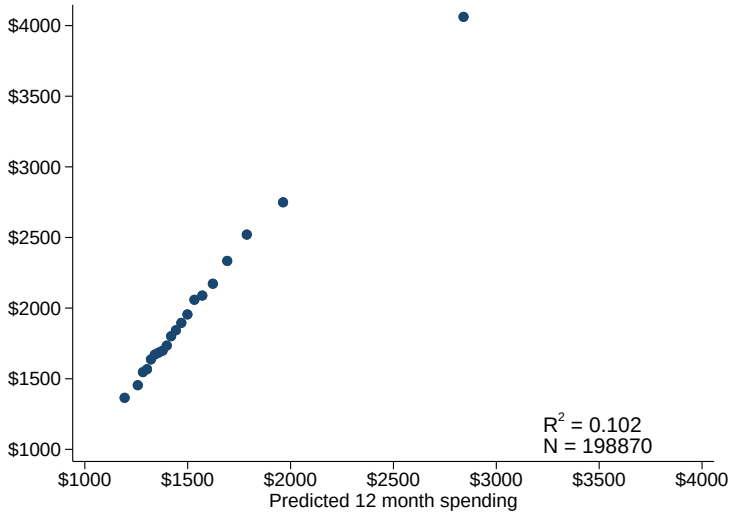
[29] Physical therapy industry representatives reportedly launched "feverish lobbying campaigns ... directed at softening" the payment changes brought on by the Balanced Budget Act (Goldstein, 1999). The American Physical Therapy Association recruited thousands of physical therapists to stage protests on Capitol Hill and organized phone-a-thons to call on Congressional representatives. (Luthra, 2018).

D Machine Learning Methodology and Results

D.1 Patient-level 12 Month PT Spending Prediction

Methodology To create a proxy for patient need, we construct a patient-level measure of predicted PT spending based on patient characteristics and utilization in the 6 months and the year *prior* to starting PT. We then apply this prediction to all 2005-2006 patients in the health analysis sample in Section 3.2 and patients who approach and attempt the cap in Section 4. The model is trained on patients who approach the cap threshold in 2004 and 2005, prior to the implementation of the therapy cap. We use gradient-boosted decision trees from the LightGBM package. The predictors are: age, sex, utilization and spending in the previous calendar year available in the MBSF Cost and Utilization file (in-office spending, Part B drug, outpatient procedure, inpatient, testing, imaging, hospice, evaluation and management, durable medical equipment, dialysis, and other), chronic conditions at the end of the previous calendar year, PT and OT spending in the previous calendar year, inpatient and SNF stays within the last 6 months (spending, number of visits, Diagnosis Related Group in most recent visit, length of stay of last visit, and days since last visit), in-office spending in the last six months, and an indicator for having an auto exception diagnosis in the last 6 months.

Figure D1: Predicted vs. Actual 12-month Spending, 2004-2005



This figure plots the predicted 12-month physical therapy spending against the actual 12-month spending in 2004-2005 from the model described in Section 2.3 and Appendix Section D. Data: 20% Medicare Carrier claims and Master Beneficiary Summary File.

Figure D1 plots the relationship between actual 12-month PT spending and predicted spending based on the model; there is a monotonic relationship between predicted and actual spending and the R^2 of the prediction is over 10 percent. The predictors with the highest feature importance are (in order of importance): Part B physician office spending in the previous year, Part B drug spending in the previous year, total in-office spending in the last 6 months, patient age, spending on tests in the previous year, spending on imaging in the previous year, other procedures spending in the previous year, durable medical equipment spending in the previous year, total outpatient spending in the last 6 months, and the number of imaging events in the previous year.

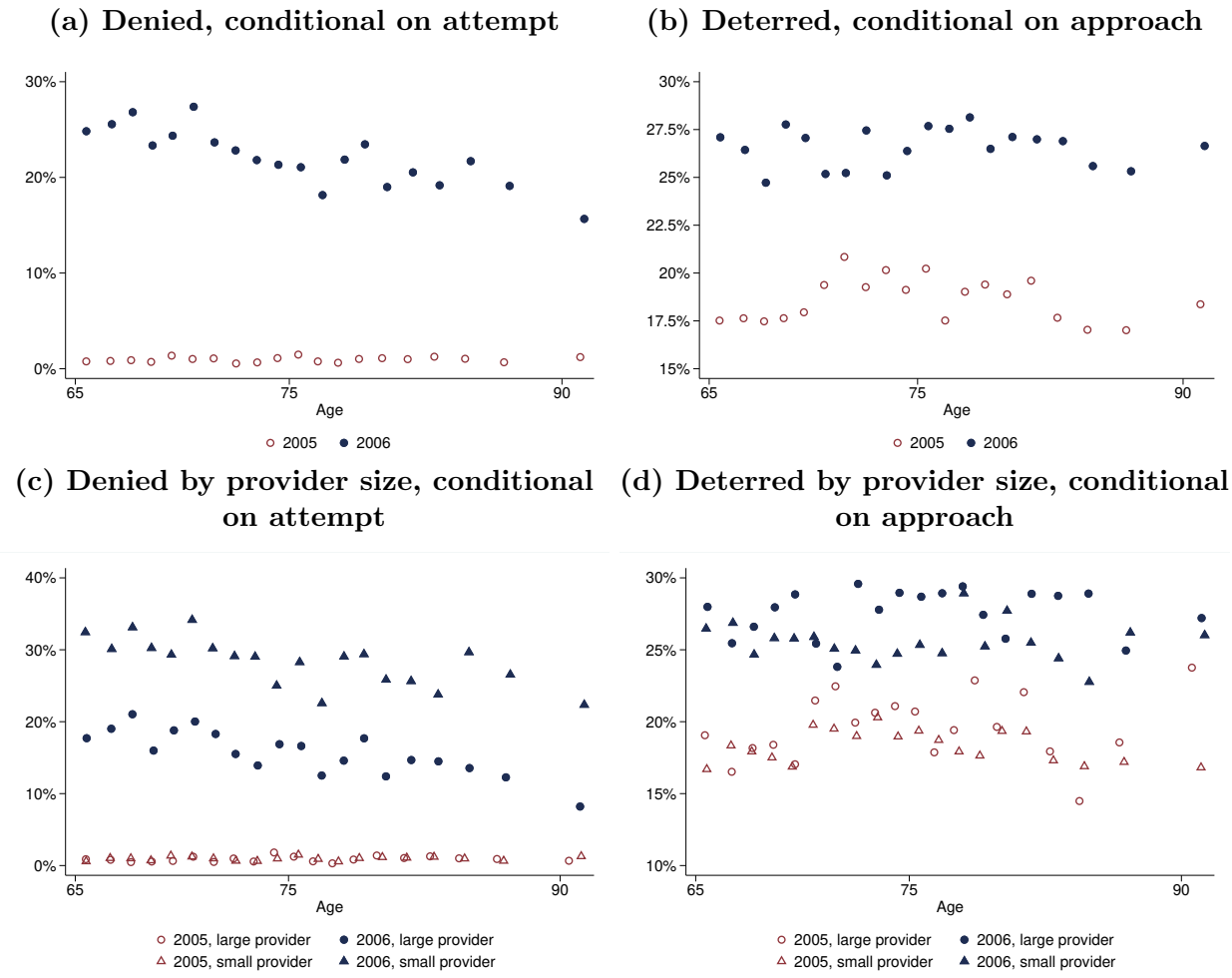
Interpretation of Predicted Spending Measure Our screening tests in Section 4 use the 12-month PT spending prediction as a proxy for a patient's true medical need for more PT. This interpretation of our predicted spending measure requires assuming that a patient's relative position in the spending distribution is informative about their true latent need, and that this ranking is stable between 2005 and 2006. Even if the overall *level* of spending in 2004 and 2005 may have been considered too high by Medicare, our test is valid as long as relatively high-spend patients in 2004-2005 have greater true PT need than relatively low-spend ones.

One potential limitation of using a predicted outcome trained on healthcare claims as our measure of clinical need is that these prediction models are known to inherit biases and reproduce human judgment errors (Mullainathan and Obermeyer, 2017). For example, lower-income populations may under-utilize care relative to their actual need because they are more sensitive to healthcare costs or due to provider bias. If prior utilization predicts greater future spending in the model, then it would underestimate true need for these populations. This is an issue inherent in any analysis using claims-based measures to proxy for patient need.

This potential for bias has several implications for the interpretation and validity of our results. For the slope-based screening test, if predicted and true need were only weakly positively correlated or completely uncorrelated with each other, then this would bias our test *against* finding screening on need. Using a biased measure of need would undermine our analysis only if the bias is large enough such that the predicted and true need were sufficiently negatively correlated with each other, which is a fairly extreme case of model misspecification. In that case, what we would classify as improved screening on need (i.e., a steepening of the slope) would actually be indicative of worsened screening. For the horizontal inequity analyses that test for a level change, this bias should generate a gap in outcomes between $C_i = 1$ patients and $C_i = 0$ patients in both the pre-reform and post-reform period. We need the magnitude and distribution of this error with respect to X_i to not change over time in order for it to not affect our coefficient of interest β_4, which is on the interaction between C_i and an indicator for $Year_{y(i)} = 2006$.

Furthermore, we validate that our results are robust to these concerns by repeating our analysis using patient age instead of predicted need in Figure D2 and Table A2. Age is plausibly correlated with true need but not prone to claims-based measurement error. We find qualitatively similar results when using patient age as our measure of need.

Figure D2: Correlations between patient age and deterrence and denial, 2005-2006



This figure plots the relationship between attempt outcomes and patient age and provider size. "Denied" is defined as the share of patients who attempt but never make it past the cap, and "deterred" is defined as the share of patients who approach the cap but do not attempt. Sample restricted to patients over the age of 65. Panels (a) and (b) plot the relationship between log age and share denied and deterred in 2005 and 2006. Panels (c) and (d) plot the same relationship, split by provider size. Provider size is defined as the total number of Medicare beneficiaries who receive regular PT by that provider in 2006-2008, and a large firm is defined as being above-median. Section 2 describes the sample definition in further detail. Data: 20% Medicare Carrier claims and Master Beneficiary Summary File.

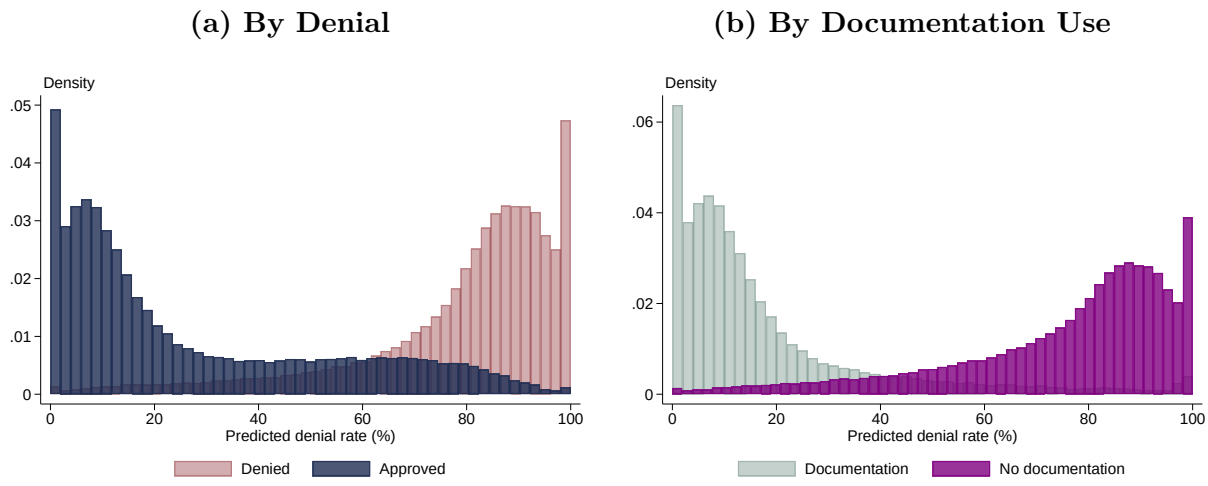
D.2 Claim-level Denial Rate Prediction

Motivation Table 3 showed that including documentation substantially increases the explanatory power of a regression model in explaining denials. However, this regression model is unable to capture the high-dimensional information contained on claims, such as procedure

codes and diagnosis codes. Thus, it may not be able to capture complex billing strategies like upcoding which could be used to increase approval rates. To explore whether strategic billing behavior could be driving approvals over time, we train a machine learning model that uses line- and claim-level information from claims associated with the attempt as well as the weeks of care leading up to the attempt.

Methodology We first use data from 2006 to train a machine learning model that predicts the likelihood of claim-level denial rate for claims aiming to exceed the cap using information from the claim. We again use gradient-boosted decision trees from the LightGBM package. The predictors include patient age, sex, ICD-9 diagnosis codes on the claim, the HCPCS procedure code on the line item, the number of units on that line item, modifier codes on the line item (including the KX modifier), number of units, modifier codes on the patient's last 5 visits, as well as the prior year and prior 6 month utilization and spending used in the 12 month PT spending prediction described above. Intuitively, this model is a probabilistic approximation to the decisions made by the Medicare contractor in deciding which claims to deny. We then apply the prediction to claims associated with attempts in 2005 to 2008, setting the KX modifier to be "on" or "off" for all lines.

Figure D3: Predictions by Denial and Documentation Status



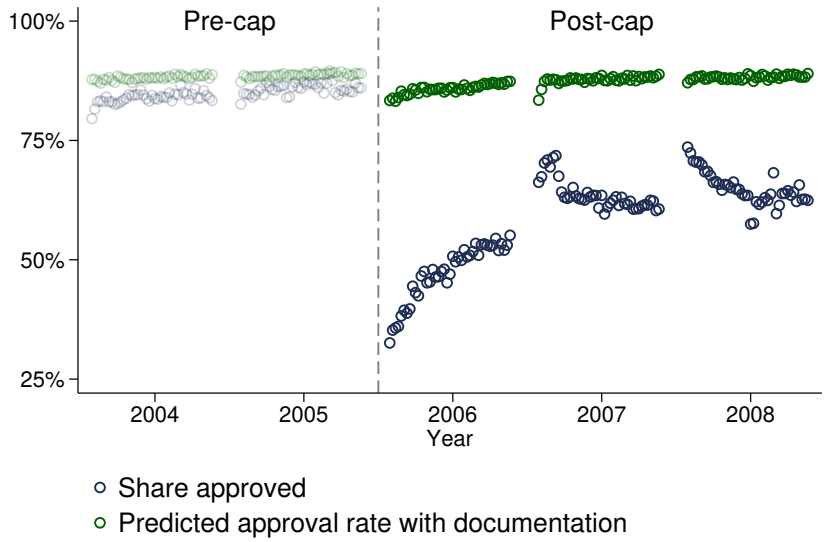
This figure plots claim-level predicted denial rates from the denial rate prediction model described in Section 5. Panel (a) plots the predicted rates for claims associated with attempts in 2006, split by whether the claim was actually denied or not. Panels (b) predicted rates, split by whether the claim had documentation or not. The prediction model is trained on 2006 claims associated with cap attempts and explanatory variables in the prediction model are discussed in Appendix Section D. Section 2.3 describes the definition of cap attempts in further detail. Data: 20% Medicare Carrier claims and Master Beneficiary Summary File.

Results Figure D3 panel (a) shows that our model predicts denials out of sample well. For denied claims, the median predicted probability of denial is 79%. We label 90 percent of denied claims as having more than 50% probability of denial. Likewise, for approved claims, the median predicted denial probability is just 16%. We label 78 percent of these approved claims as having less than 50% probability of denial.

While the model incorporates many more variables than the regression in Table 3, again we find that one factor has outsize predictive influence: documentation. Figure D3 panel (b) shows that even though the machine learning model was trained using a host of patient- and claim-level characteristics, the documentation indicator alone explains much of the variation in the predicted denial rate. Indeed, this is because documentation explains a large amount of variation in the actual denial rate: in 2006, just 37% of claims using documentation were denied, while 75% of claims without documentation were denied.

Finally, we investigate whether the change in approval rates over time could reflect changes in strategic billing behavior. In order to capture billing behavior that is independent of documentation usage, we use our model to create a predicted approval rate under the assumption that *every attempt uses documentation*. The time series of this modified predicted approval rate in Figure D4 shows that changes in billing behavior besides documentation are unable to account for most of the increase in the approval rate over time: if all claims were coded as having documentation, the approval rates would have largely persisted at their pre-cap levels.

Figure D4: Time Series Variation in Documentation and Predicted Approval Rate on Cap Attempts



This figure plots the share of attempt weeks with no denials ("Share approved") and predicted approval rate based on claim characteristics, assuming documentation ("Predicted approval rate with documentation"). The predicted approval rate is derived from the line-level denial rate prediction described in Section D on claims associated with attempts, where the KX modifier indicator for documentation use is turned on for all claims. 20% Medicare Carrier claims.

E Construction of Patient Health Measures

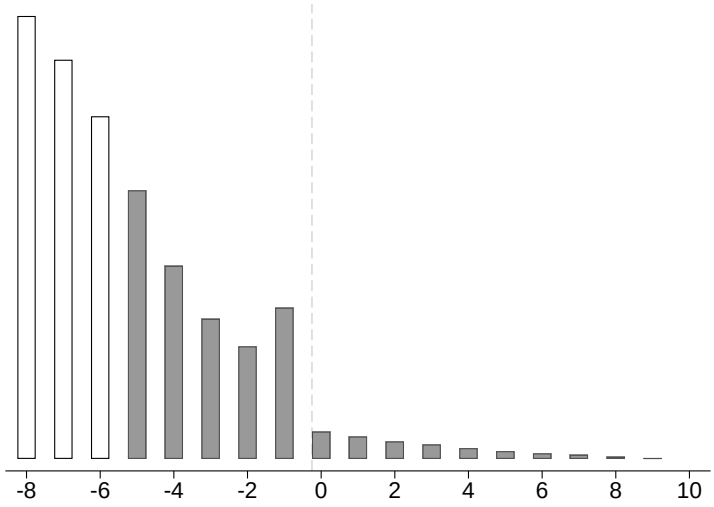
Injury diagnoses were identified as claims within 90 days of the PT start date in MEDPAR (inpatient hospital stay), Outpatient, and Carrier files an ICD-9 code starting with 800-899. Pain management procedures were identified in the Outpatient and Carrier files based on HCPCS codes. Orthopedic surgeries were identified in the Outpatient files using based on HCPCS codes, respectively. Crosswalks to HCPCS codes for pain procedures and orthopedic surgeries were created in consultation with a clinical expert (available upon request). Emergency department visits were identified as inpatient stays in MEDPAR with positive ED charges or Outpatient claims with a Revenue Center Code between 450 and 459 or equal to 0981. Hospital stays were identified in MEDPAR as claims with Short Stay/Long Stay/SNF Provider Indicator Code "S" or "L," and skilled nursing stays were identified as claims with Short Stay/Long Stay/SNF Provider Indicator Code "N." Opioid prescriptions were identified in the Part D file as prescriptions with Product Service ID/National Drug Service Code via a crosswalk to Anatomical Therapeutic Chemical codes for opioids, in consultation with a clinical expert (available upon request).

F Patient Health Effects: Difference-in-Difference Strategy

Sample and Identification Strategy In addition to the identification strategy in Section 3.2 which leverages within-year differences in spending depending on when a patient starts PT, we deploy an alternative empirical strategy to assess the patient health effects of the cap. This empirical strategy is in the spirit of the estimator used in Diamond and Persson (2016), which proposes a method to estimate the causal effects of bunching by comparing outcomes for individuals inside and outside of a "manipulation region" around a discontinuity. In our case, we compare average health outcomes among patients whose end-of-year spending could have plausibly have been reduced by the cap, who serve as the treated group, and patients whose spending is too low to have been affected by the cap, who serve as the control group. Figure F1 illustrates how we define the two groups. The treated group includes any patients who are over the cap or within 5 weeks of the cap, while the control group includes all patients over 5 weeks under the cap.[30] The underlying assumption is that the control group does not include any "bunchers" whose spending was at risk of being affected by the cap.

[30]Because we convert a patient's end-of-year spending into "weeks from cap" using the maximum of their 5-week rolling average spending *or* the sample average weekly spending, this left-censors the "weeks from cap" measure at -8. The -8 week bin can be interpreted as patients who appear to end the year 8 or more weeks below the cap. We also restrict to patients who are at most 48 weeks to the right of the cap to exclude patients who are implausibly far from the cap. We also restrict to patients who end their PT outside of the first and last 4 weeks of the year.

Figure F1: Patient Health Outcomes DD Identification Strategy



This figure illustrates the treatment (gray) and control (white) group assignment for the difference-in-difference patient health identification strategy described in Section F.

If the reductions in spending from the cap led to worsened patient health, then once the cap is in place, we would expect the treated group's *average* health to fall relative to the control group. To see this, consider a patient who *would* have ended up above the cap, but instead bunches under the cap once it is in place. This patient contributes to the treated group's average both before and after the reform. Thus, if reducing spending harmed this patient, we would expect the average health for the treated group should fall. In contrast, the control group is comprised of patients who would remain far below the cap regardless of whether the cap is in place. We do not have to know *which* patients below the cap are there because of the cap, but rather just that some share of them would have counterfactually been above the cap.

After constructing the treatment and control groups, we use a difference-in-difference strategy to compare how their health outcomes evolve before and after the 2006 cap. Interpreting the estimates from this difference-in-difference as causal requires making two assumptions. The first is the standard parallel trends assumption: the health trends for each group would be parallel in the absence of the reform. We will verify this by looking for evidence of pre-trends in an event study. The second assumption is that the composition of patients in each group is unaffected by the cap. In other words, none of the "bunchers" reduced their spending so much that they ended up over 5 weeks away from the cap. Figure 3 shows that there is no statistically significant difference in the 2005 and 2006 distributions past 4 weeks from the cap. Additionally, the lack of an extensive margin response in Figure B4 suggests

that the cap did not affect the composition of patients who seek PT. Furthermore, in our main specification we directly control for several observable patient characteristics.

Results We present the results in the form of a yearly event study. The specification for patient i receiving care in $Year_i$ that ends in calendar week $LastWeek_i \in [1, 52]$ is:

$$Y_i = \beta_0 + \beta_1 Treated_i + \sum_{\tau=2004}^{2008} \beta_{2\tau}\ Treated_i \times 1(Year_i = \tau) + LastWeek_i + \varepsilon_i. \tag{6}$$

The pooled specification is:

$$Y_i = \beta_0 + \beta_1 Treated_i + \beta_2\ Post_i + LastWeek_i + \varepsilon_i, \tag{7}$$

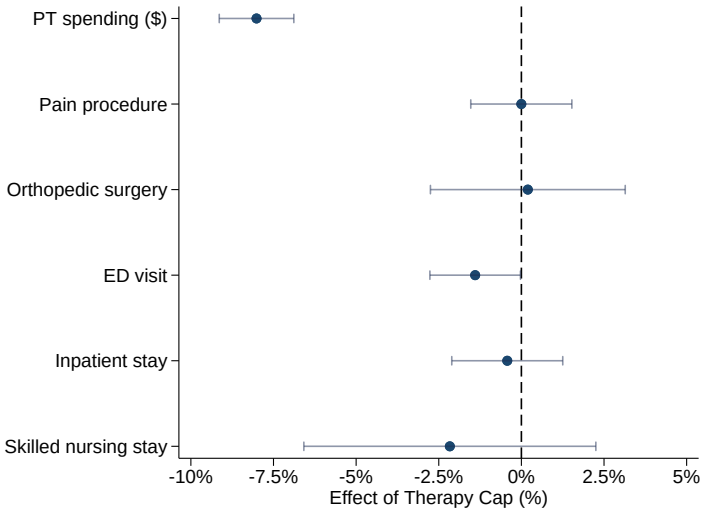
where $Post_i$ is an indicator for years after 2006. The standard errors are clustered at the patient level and the omitted year is 2005. Figure F2 plots the $\beta_{2\tau}$ estimates from equation (6) and Figure F3 plots the coefficients the pooled specification in equation (7). Figure F2 panel (a) confirms a "first stage"—the treated group saw a sizable reduction in PT spending relative to the control group. Turning to health outcomes, consistent with the null results found in Section 3.2, we find no evidence of increases in (b) the usage pain management procedures, (c) orthopedic surgeries, (d) hospital stays, or (e) skilled nursing home stays.[31]

[31] Given that the Medicare Part D program for prescription medication began in 2006, we cannot use this difference-in-difference strategy to study opioid prescriptions.

Figure F2: DD Design: Spending and Health Outcomes within 12 Months of First PT

This figure plots the coefficients from the health outcomes regression in equation (6). Panel (a) plots the coefficients on log 12-month PT spending. All outcomes are all measured within 12 months of first PT session. Panel (b) plots the coefficients on an indicator for pain management procedures, panel (c) plots the coefficients on an indicator for orthopedic surgery, panel (d) plots the coefficients on an indicator for an emergency department visit, panel (e) plots the coefficients on an indicator for a hospital stay, and panel (f) plots the coefficients on an indicator for a skilled nursing facility stay. The left-side y-axis denotes the effect in terms of percentage points and the right-side y-axis denotes the effect in terms of percent of the control group average. The treated group comprises patients who end the year within 5 weeks away from the cap or above the cap, and the control group compromises patients who end the year over 5 weeks below the cap. Section 2.3 provides additional detail on how "weeks from the cap" is defined. Data: 20% Medicare Carrier claims and Master Beneficiary Summary Files.

Figure F3: DD Design: Spending and Health Outcomes within 12 Months of First PT



This figure plots the coefficients from the health outcomes regression in equation (7). All outcomes are all measured within 12 months of first PT session in a given year. The treated group comprises patients who end the year within 5 weeks away from the cap or above the cap, and the control group compromises patients who end the year over 5 weeks below the cap. Section 2.3 provides additional detail on how "weeks from the cap" is defined. Data: 20% Medicare Carrier claims and Master Beneficiary Summary Files.

G Conceptual Framework for Characterizing Screening

This section develops a potential outcomes framework for assessing whether the Medicare therapy cap screens patients on the basis of need. We then show that, under a linearity assumption, the reduced-form slope test in Section 4 is equivalent to the standard condition for screening.

Let $t \in \{0, 1\}$ be an indicator for whether the cap is in effect, and define $D_i(t)$ and $Q_i(t)$ for patient i approaching the cap to be the potential outcomes of deterrence and denial, respectively, if she faced treatment t. We then define stopping at the cap as facing either a deterrence or a denial: $S_i(t) = D_i(t) + (1 - D_i(t))Q_i(t)$. We assume monotonicity and nondegeneracy in the potential outcomes so that the cap weakly increases deterrence and denials and generates a positive mass of compliers in each channel. Under monotonicity, $\Delta D_i \equiv D_i(1) - D_i(0)$ is an indicator that characterizes whether i is a "deterrence complier"—i.e., they would only be deterred when the cap is in place— $\Delta Q_i \equiv Q_i(1) - Q_i(0)$ characterizes whether i is "denial complier"—i.e., they would only have a submitted claim be denied when the cap is in place, and $\Delta S_i \equiv S_i(1) - S_i(0)$ characterizes whether i is a "cap complier"—i.e., they are stopped by the cap. Finally, we let X_i be a continuous measure of patient i's medical need for PT care.

The recent literature on the targeting properties of ordeals (Alatas et al., 2016; Deshpande and Li, 2019; Finkelstein and Notowidigdo, 2019) has typically defined screening based on the extent to which compliers have lower-than-average need. Definitions G.1 and G.2 follow this convention.

Definition G.1 (Deterrence on Need)**.** *The extent to which the cap screens on need through deterrence—i.e., "deters on need"—is characterized by the magnitude of:*

$$\mathbb{E}[X_i \mid \Delta D_i = 1] - \mathbb{E}[X_i],$$

where a negative value indicates that the cap deters low-need care.

Definition G.2 (Denial on Need)**.** *The extent to which the cap screens on need through denials—i.e., "denies on need"—is characterized by the magnitude of:*

$$\mathbb{E}[X_i \mid \Delta Q_i = 1, D_i(1) = 0] - \mathbb{E}[X_i | D_i(1) = 0],$$

where the condition $D_i(1) = 0$ restricts to individuals who would attempt to bill (i.e., are not deterred) when facing the cap. A negative value indicates that the cap denies low-need care.

Definition G.3 (Screening on Need Overall)**.** *The extent to which the cap screens on need overall is characterized by the magnitude of:*

$$\mathbb{E}[X_i \mid \Delta S_i = 1] - \mathbb{E}[X_i],$$

where a negative value indicates that the cap screens out low-need care.

In our analysis in Section 4, we run the following regressions on observed data:

$$D_i = \beta_1^D X_i + \beta_2^D X_i \times 1(t(w(i)) = 1) + \gamma_{w(i)}^D + \varepsilon_i^D, \tag{8}$$

$$Q_i = \beta_1^Q X_i + \beta_2^Q X_i \times 1(t(w(i)) = 1) + \gamma_{w(i)}^Q + \varepsilon_i^Q, \tag{9}$$

$$S_i = \beta_1^S X_i + \beta_2^S X_i \times 1(t(w(i)) = 1) + \gamma_{w(i)}^S + \varepsilon_i^S, \tag{10}$$

where $w(i)$ is the week of attempt or approach and $t(w(i))$ denotes the treatment status for i implied by the timing of their approach or attempt (i.e., whether it occurred in 2005 or 2006). Regressions (8) and (10) are estimated on patients who approach the cap, while (9) is estimated on the subpopulation that attempts to bill at the cap.

By imposing the following two assumptions, we can map the estimates from our regressions to definitions G.1 and G.2 through the subsequent two propositions.

Assumption 1. *The distribution of* $(D_i(1), D_i(0), Q_i(1), Q_i(0), X_i)$ *is independent of* $w(i)$*. The timing of when a patient approaches or attempts to go over the cap is independent of their need or potential outcomes.*

Assumption 2. *For each* $t \in \{0, 1\}$*, there exist constants* α_t, θ_t *such that*

$$\mathbb{E}[Q_i(t) \mid X_i, D_i(0), D_i(1)] = \mathbb{E}[Q_i(t) \mid X_i] = \alpha_t + \theta_t X_i.$$

the conditional expectation of the potential outcomes for denials is linear in need and, conditional on need, does not vary with the deterrence potential outcomes.

Proposition 1. *The magnitude of deterrence on need is the following scaling of the difference in slopes:*

$$E[X_i|\Delta D_i = 1] - E[X_i] = \beta_2^D \frac{\text{Var}(X_i)}{E\left[\Delta D_i\right]} \tag{11}$$

and the magnitude of screening on need overall is the following scaling:

$$E[X_i|\Delta S_i = 1] - E[X_i] = \beta_2^S \frac{\text{Var}(X_i)}{E\left[\Delta S_i\right]} \tag{12}$$

Proof.

$$
\begin{aligned}
E[X_i|\Delta D_i = 1] - E[X_i] &= \frac{\mathbb{E}[X_i \Delta D_i]}{\mathbb{E}[\Delta D_i]} - \frac{E[X_i]E[\Delta D_i]}{E[\Delta D_i]} \\
&= \frac{\text{Cov}(X_i, \Delta D_i)}{E[\Delta D_i]} && (13) \\
&= \frac{\text{Cov}(D_i(1), X_i) - \text{Cov}(D_i(0), X_i)}{E[\Delta D_i]} && (14) \\
&= \left(\frac{\text{Cov}(D_i(1), X_i)}{\text{Var}(X_i)} - \frac{\text{Cov}(D_i(0), X_i)}{\text{Var}(X_i)} \right) \frac{\text{Var}(X_i)}{E[\Delta D_i]} && (15) \\
&= \left(\underbrace{\frac{\text{Cov}(D_i, X_i \mid t = 1)}{\text{Var}(X_i \mid t = 1)}}_{\text{Post-cap slope}} - \underbrace{\frac{\text{Cov}(D_i, X_i \mid t = 0)}{\text{Var}(X_i \mid t = 0)}}_{\text{Pre-cap slope}} \right) \frac{\text{Var}(X_i)}{E[\Delta D_i]} && (16) \\
&= \beta_2^D \frac{\text{Var}(X_i)}{E[\Delta D_i]}, && (17)
\end{aligned}
$$

where we use Assumption 1 to translate from potential outcomes (equation 15) to observed outcomes (equation 16). The proof for screening on need overall follows the same logic. □

Proposition 2. *The magnitude of denial on need is the following scaling of the difference in slopes:*

$$
E[X_i \mid \Delta Q_i = 1, D_i(1) = 0] - E[X_i \mid D_i(1) = 0] = \beta_2^Q \frac{\text{Var}(X_i \mid D_i(1) = 0)}{E\left[\Delta Q_i \mid D_i(1) = 0\right]} \tag{18}
$$

Proof. For ease of notation, we define E_Q, Cov_Q, and Var_Q to imply integrating while conditioning on $D_i(1) = 0$ (i.e., making an attempt when the cap is in place). Analogously to

for deterrence:

$$
\begin{aligned}
E_Q[X_i|\Delta Q_i = 1] - E_Q[X_i] &= \frac{E_Q[X_i \Delta Q_i]}{E_Q[\Delta Q_i]} - \frac{E_Q[X_i]E_Q[\Delta Q_i]}{E_Q[\Delta Q_i]} \\
&= \frac{\text{Cov}_Q(X_i, \Delta Q_i)}{E_Q[\Delta Q_i]} && (19) \\
&= \frac{\text{Cov}_Q(Q_i(1), X_i) - \text{Cov}_Q(Q_i(0), X_i)}{E_Q[\Delta Q_i]} && (20) \\
&= \left(\frac{\text{Cov}_Q(Q_i(1), X_i)}{\text{Var}_Q(X_i)} - \frac{\text{Cov}_Q(Q_i(0), X_i)}{\text{Var}_Q(X_i)} \right) \frac{\text{Var}_Q(X_i)}{E_Q[\Delta Q_i]} && (21) \\
&= \left(\underbrace{\frac{\text{Cov}_Q(Q_i, X_i \mid t = 1)}{\text{Var}_Q(X_i \mid t = 1)}}_{\text{Post-cap slope}} - \frac{\text{Cov}_Q(Q_i, X_i \mid t = 0)}{\text{Var}_Q(X_i \mid t = 0)} \right) \frac{\text{Var}_Q(X_i)}{E_Q[\Delta Q_i]} && (22) \\
&= \left(\underbrace{\frac{\text{Cov}_Q(Q_i, X_i \mid t = 1)}{\text{Var}_Q(X_i \mid t = 1)}}_{\text{Post-cap slope}} - \underbrace{\frac{\text{Cov}(Q_i, X_i \mid t = 0, D_i(0) = 0)}{\text{Var}(X_i \mid t = 0, D_i(0) = 0)}}_{\text{Pre-cap slope}} \right) \frac{\text{Var}_Q(X_i)}{E_Q[\Delta Q_i]} && (23) \\
&= \beta_2^Q \frac{\text{Var}_Q(X_i)}{E_Q[\Delta Q_i]}, && (24)
\end{aligned}
$$

where again we use Assumption 1 to translate from potential outcomes to observed outcomes. However, we require additional Assumption 2 in order to apply the equality:

$$
\frac{\text{Cov}_Q(Q_i, X_i \mid t = 0)}{\text{Var}_Q(X_i \mid t = 0)} = \frac{\text{Cov}(Q_i, X_i \mid t = 0, D_i(1) = 0)}{\text{Var}(X_i \mid t = 0, D_i(1) = 0)} = \frac{\text{Cov}(Q_i, X_i \mid t = 0, D_i(0) = 0)}{\text{Var}(X_i \mid t = 0, D_i(0) = 0)},
$$

where the first equality follows from Assumption 1. In particular, Assumption 2 ensures that the slope estimated pre-cap is the same whether conditioning on the known population for which $D_i(0) = 0$ or the unknown population for which $D_i(1) = 0$.

□

Corollary 0.1. *β_2^D, β_2^Q, and β_2^S have the same signs as $E[X_i|\Delta D_i = 1] - E[X_i]$, $E[X_i \mid \Delta Q_i = 1, D_i(1) = 0] - E[X_i \mid D_i(1) = 0]$, and $E[X_i|\Delta S_i = 1] - E[X_i]$, respectively.*

Proof. $\text{Var}(X_i), \text{Var}_Q(X_i) \geq 0$ trivially, and $E[\Delta D_i], E_Q[\Delta Q_i], E[\Delta S_i] > 0$ by monotonicity.

□

H Sharp Evidence of Learning-By-Doing

To complement our decomposition evidence in Section 5 on provider learning-by-doing, in this section we look at how provider behavior evolves around events which should be associated with learning. For each provider, we identify the point at which they seem to "learn" how to avoid a denial by looking for the first time that the provider successfully reverses a denial on a previous attempt with a patient. In particular, we look for the weeks in which a provider makes an approved attempt *after* receiving a denial in an attempt with the same patient in a prior week.

We use a stacked event study method, following Cengiz et al. (2019). Each stack is centered around a given provider's learning event with a focal patient — the "treatment" — and we look at that provider's five attempt weeks before and after their event, with patients *other* than the focal one. We use attempt weeks instead of calendar weeks so as to only capture weeks in which the provider makes at least one attempt. These observations form the treated group within each stack. The control group within each stack are "clean controls" comprised of *other* providers who make attempts in the same calendar week as the learning event, but who do not have a change in treatment status in the 5 attempt weeks before or after the focal week. In other words, they are either not treated in the entire period (because they are not-yet-treated or they are never-treated) or treated in the entire period. A given attempt can appear in the control group for multiple stacks, but can only be the learning event that defines treatment in one stack.

The regression specification for patient i's attempt within stacked group g is:

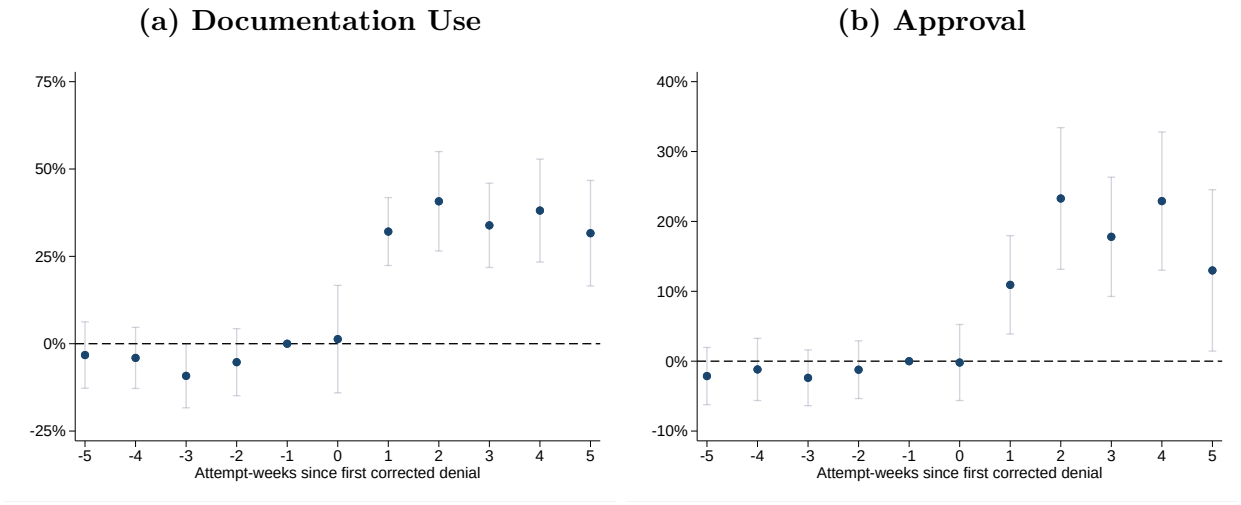
$$Y_{igt} = \sum_{a(i,g)=-5}^{5} 1(RelWeek_{a(i,g)} = a) \times Treat_{j(i),g} + \underbrace{\alpha_{j(i),g}}_{\text{Provider-Group FE}} + \underbrace{RelWeek_{a(i,g),g}}_{\substack{\text{Week Relative to} \\ \text{Attempt-Group FE}}} + \underbrace{Week_{t,g}}_{\substack{\text{Calendar Week-} \\ \text{Group FE}}} + \underbrace{\beta X_i}_{\text{Patient Controls}} + \varepsilon_{igt}. \tag{25}$$

The outcome of interest, Y_{igt}, is either documentation use for attempts or approvals on attempts in calendar week t. $Treat_{j,g}$ is an indicator variable for whether the attempt with patient i is associated with provider j's "learning event" and $\alpha_{j,g}$ is a provider-group fixed effect. $RelWeek_{a,g}$ is the attempt week (between -5 and 5) *relative* to the learning event that defines group g, and $Week_{t,g}$ is the calendar week associated with patient i's attempt (interacted with group g). These can be separately identified because not all providers make an attempt in every week. We also control for X_i, the predicted need associated with the

patient. Results are clustered at the provider-level to account for repeated observations across stacks.

Figure H1 plots the coefficients from equation (25): panel (a) shows the results for documentation use during the attempt week, and panel (b) shows the results for whether the attempt was approved (i.e., not denied). Both sets of results show that prior to the learning event, treated providers were not on a differential trend relative to control providers. However, after the provider corrects their first denial, they are consistently more likely to include documentation on future attempts and more likely to be approved. These results suggest sharp learning-by-doing within a provider—once they successfully reverse their first denial, they change their billing behavior to ensure future attempts are approved.

Figure H1: Outcomes on Attempts Around Provider Learning Event, 2006



Notes: This figure plots the coefficients from the stacked event study specification around a provider's "learning event" in equation (25). The outcome variables are (a) the share of attempts with documentation and (b) the share of attempts approved (i.e., not denied). The learning event is defined as a week in which a provider makes an approved attempt with a patient after receiving a denied attempt in a previous week with the same patient in a prior week. The sample is of attempts with patients *other* than the one associated with the learning event. An attempt-week is a week in which the provider made at least one attempt, as defined in Section 2.3. Each learning event is grouped and compared to other providers who also made attempts in the same calendar week, and the groups are stacked together (Cengiz et al., 2019). The specification includes controls for patient predicted PT spending, provider-group, and week-group fixed effects, and is clustered at the provider and group levels. Section 2.3 describes the definition of cap attempts in further detail. Data: 20% Medicare Carrier claims.