Screening Through Spending Limits: Evidence from the Medicare Therapy Cap*

Ashvin Gandhi
UCLA & NBER

Maggie Shi
UChicago Harris & NBER

February 13, 2025

Abstract

Governments and firms frequently use spending limits to control costs, but still allowing case-by-case exceptions supported by documentation. This paper studies a Medicare policy which capped per-patient physical therapy spending and added a documentation requirement to bypass the limit. The cap reduced spending by 8 percent and improved targeting on clinical need without harming patient health, but also introduced horizontal inequity conditional on need. The targeting improvements arise from Medicare being discerning in granting exceptions, rather than improved screening by providers. Conditional on need, lower-income and minority patients are substantially more likely to be screened out by the cap, as they tend to go to smaller providers who have poorer documentation practices.

*We thank Zhijian Li 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, Josh Gottlieb, Tim Layton, Jetson Leder-Luis, Riley League, Ryan McDevitt, 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, American Health Econometrics Workshop, the Southeastern Health Economics Study Group, Penn Leonard Davis Institute, UChicago Harris, and Bowdoin College for helpful comments. The authors gratefully acknowledge support from the National Institute on Aging (#T32-AG000186) and Arnold Ventures. All errors are our own.

1 Introduction

Governments and firms frequently delegate spending decisions to agents. But if these agents have different objectives than the principal, they may choose to spend in ways the principal would consider wasteful. One way to address this is to impose a "soft" spending limit on the agent — a policy that restricts spending above a stated amount, but still allows for exceptions on a case-by-case basis. Under these policies, agents must submit documentation to request approval from the principal to go past the limit.

Policies like spending limits employ documentation to screen out waste in two ways: not only does the principal use the embedded information to decide whether to *deny* the request, but the ordeal of the submission can itself *deter* agents from even attempting to make low-value requests (Nichols and Zeckhauser, 1982). However, documentation requirements can introduce horizontal inequity if they favor agents with greater administrative capacity — all else equal, agents with lower documentation costs or who make fewer mistakes are less likely to be deterred or denied. While many screening mechanisms employ documentation requirements to target waste,[1] there has been little empirical work unpacking how exactly they screen and the equity-efficiency tradeoff they impose.

This paper studies these questions in the context of a spending limit imposed by insurers onto healthcare providers. Insurers delegate care decisions to providers, who deliver care and then submit claims for reimbursement. Insurers mitigate overspending by setting per-patient spending limits on certain services and requiring providers to submit documentation to justify bypassing them. The required documentation can be complex and includes visit notes, test results, imaging reports, evidence of previous treatments, and referral letters. Spending limits are common for repeated services where concerns about waste are mostly on the intensive margin, like therapy visits – the focus of this paper – as well as prescription drug refills and extended nursing home or hospice stays.

This setting – a soft spending limit imposed on a healthcare service – is well-suited for

[1]For example, government agencies impose procurement thresholds where purchases above a certain limit require documentation to justify that the price is reasonable and must be approved by a higher authority (Code of Federal Regulations, 2020). To receive disability benefits, individuals must submit medical evidence supporting their disability claim to an examiner, who can deny the application if they disagree (Social Security Administration, 2008). When making an insurance claim, individuals must submit documentation of damages and repair costs, but the eventual decision is made by an adjuster who can revise or reject it.

understanding screening via documentation. First, the transaction-level detail in healthcare claims data allows us to distinguish between cases screened out through deterrence – those that stop at the limit without submitting a claim to go over – and cases screened out through denials – those who submit a claim but are denied. Second, the limit is a sharp threshold, so we can precisely characterize who is screened out by focusing on the "bunchers" immediately below the threshold. Third, we can use patient health outcomes as an objective measure to test whether the limit successfully screened out wasteful spending.

In particular, we study a spending limit for physical therapy (PT) within the Medicare program, which insures the elderly and disabled in the United States. The "therapy cap" was an annual, per-patient spending limit of $1,740, or about 11 weeks of PT care. Providers could bypass the limit by requesting an exception for a patient, but the request had to be supported by documentation. Medicare then approved or denied these requests based on their assessment on the patient's medical necessity.[2] Across most of our analyses, we distinguish between the deterrence channel, which is captured by whether the provider submits a claim that bypasses the limit, from the denial channel, which is captured by whether Medicare denies such a claim.

We first quantify the total savings from the cap using a difference-in-bunching estimator that compares the PT spending distributions before and after the 2006 introduction of the cap. We find that overall, the therapy cap reduced spending by 8 percent.[3] 58 percent of these savings are due to Medicare denying attempts to go past the cap while the remaining 42 percent stems from providers being deterred from making these attempts.

We next examine the screening properties of these spending reductions using two separate approaches. The first approach considers the overall reduction in spending and looks for whether this directly affected patient health outcomes. If the savings were poorly targeted and reduced *necessary* care, then patient health should worsen once the cap is in place. In contrast, a lack of patient health effects would imply that the limit mostly curtailed waste. We leverage the fact that each patient's cap resets at the beginning of the calendar year,

[2]This determination is technically made by contracted firms called Medicare Administrative Contractors (League, 2022) who administer the Medicare program.

[3]To put this magnitude into perspective, we contrast it with a similar analysis on a one-year "hard" cap in 1999 that restricted *all* spending over the limit, where we find a 14 percent reduction in spending.

making it more binding for patients who start PT earlier in the year. Comparing patients who start earlier versus later in the year, before and after the cap is introduced, we find precise null effects on relevant health outcomes like hospitalizations, emergency department visits, opioid usage, pain procedures, and orthopedic surgeries.

Our second approach then distinguishes between the screening properties of the deterrence and denial channels separately. We focus on patients close to the cap, and our test for improved targeting is whether the likelihood of being deterred or denied at the cap becomes more correlated with need, relative to a placebo cap in a pre-reform year. Our proxy for patient need is expected 12-month PT spending, as predicted using all of the patient's utilization and diagnoses in the months before they start PT.

Looking at the screening properties of the deterrence and denial channels separately, we find that the targeting improvements appear to be driven entirely by denials. Medicare denials become much more correlated with need after the reform, but there is no change in the correlation between provider deterrence and need. While providers made fewer attempts overall, the spending reductions associated with provider deterrence were uniform across the distribution of patient need, indicating that that the ordeal imposed on providers is a fairly blunt screening tool. Instead, all of the targeting improvements can be attributed to the regulator's actions.

We then turn to the question of horizontal inequity: did the documentation requirement also "screen" on provider administrative capacity? Larger providers are known to have greater administrative capacity, so we repeat our targeting analysis separately for providers with above- and below-median patient counts (League, 2022; Dunn et al., 2024). There is a large disparity by provider size: at the median level of need, a patient approaching the cap who goes to a small provider is 22 percent (9.6 percentage points) more likely to get past it than one who goes to a small provider. This gap is driven entirely by denials, as there is no gap in provider willingness to attempt conditional on need. It emerges only once the cap is in place, indicating it is the directly caused by the cap.

Because in our setting smaller providers tend to see minority and low-income patients, this disparity by provider size translates into meaningful horizontal inequity in who gets past the cap across demographic groups. At the median level of need for patients who approach

the cap, the denial rate for non-white patients is 11 percent (5.2 percentage points) higher than it is for white patients. Similarly, the denial rate for patients also enrolled in Medicaid is 9 percent (4.3 percentage points) higher than for those who are not, and the denial rate for patients receiving the Part D Low-Income Subsidy is 9 percent (4.5 percentage points) higher than those who do not. These gaps across demographic groups are driven entirely by the identity of a patient's provider – the gap disappears entirely once we account for provider fixed effects.

Finally, we investigate the mechanisms driving this disparity in denials. We demonstrate that denials are almost entirely driven by whether a provider correctly uses documentation when making a request to bypass the cap. Larger providers are much more likely to use documentation and thus have lower denial rates. A large part of the size advantage comes from learning-by-doing about documentation usage – the more encounters a provider has with the therapy cap, the more likely they are to use documentation. Large providers simply have more chances to learn from experience than smaller providers.

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; Ida et al., 2022; Brot-Goldberg et al., 2023). Many screening mechanisms feature a deterrence stage where agents can "self-target," which is then followed by a denial stage where the regulator makes the final decision. Our paper is the first to distinguish between both channels within one screening mechanism. Making this distinction is crucial to understanding the mechanisms driving the overall screening effects of these mechanisms. We find that even in a setting where agents have substantial expertise, the deterrence channel is a blunt tool that reduces spending but not in a targeted way. The denial channel is more fine-tuned in its ability to target on need, but also introduces a sizable advantage to agents with better administrative capacity.

We also contribute to the literature on the drivers of disparities in the takeup of public programs. We explore a largely unexamined tradeoff of increasing policy complexity: programs that require substantial paperwork can introduce horizontal inequity when administrative capacity is heterogeneous. In our setting, variation across healthcare providers translates into meaningful disparities by race and income among patients with the *same*

underlying need. The role providers play in our setting as intermediaries between patients and Medicare is 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), who all submit paperwork that determine their client's access to public programs. Furthermore, we show that if there is learning-by-doing about new regulation, larger intermediaries will naturally have an advantage simply because they have more opportunities to learn. The large disparities we find indicates that these differences across intermediaries should be taken into consideration when policymakers are designing screening mechanisms and paperwork requirements.

Finally, we contribute to the literature on policies aimed at reducing wasteful health-care spending. This includes tools like cost-sharing (Manning et al., 1987; Chandra et al., 2010; Lieber, 2017; Brot-Goldberg et al., 2017), prior authorization (Dillender, 2018; Brot-Goldberg et al., 2023; Eliason et al., 2021), audits or denials (Agafiev Macambira et al., 2022; Shi, 2024), and anti-fraud enforcement (Howard and McCarthy, 2021; O'Malley et al., 2021; Leder-Luis, 2023). The focus of this literature has primarily been on tools aimed at the first unit or the entire episode of care, but they might not be well-suited to address waste on the intensive margin. We contribute to this literature by studying spending limits, which are intended specifically to mitigate waste for services with high initial benefit but diminishing marginal returns.

The rest of the paper proceeds as follows. Section 2 discusses the policy context and places it within a conceptual framework, and also describes the data we use and measures we construct. Section 3 presents estimates of the overall effects on savings and patient health. Section 4 characterizes screening properties of the cap at different levels of need as well as horizontal inequity among patients with the same need. Section 5 explores the mechanisms driving the horizontal inequity, and Section 6 concludes.

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 (7% of all traditional Medicare beneficiaries). Between 2012 and 2017, 16 million beneficiaries used outpatient physical therapy at least once. In this section, we first describe outpatient therapy and its coverage by Medicare, discuss the "therapy caps" implemented to reduce spending on outpatient therapy, and discuss the data used in our analysis.

Medicare beneficiaries receive can three types of therapy in the outpatient setting: physical therapy (PT), occupational therapy (OT),[4] and speech-language pathology (SLP).[5] PT accounts for over three-quarters of total outpatient therapy spending. As its name suggests, physical therapy includes the diagnosis, management, and prevention of physical dysfunction and pain. Within the Medicare population, the most common procedures associated with PT are include therapeutic exercises, manual therapy, and electronic stimulation. Table A1 lists the top PT procedure codes in our analysis sample. The most common diagnoses associated with PT are low back pain, shoulder pain, and lower leg pain.

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.[6] The plan of care includes a diagnosis, long-term treatment goals, and the type, quantity, duration, and frequency of therapy services. PT services are primarily provided in a private practices (34% of spending percent of 2018 spending), nursing facilities (38%), and hospitals (15%) (American Physical

[4]Occupational therapy can resemble physical therapy—e.g., OT frequently also frequently entails therapeutic exercises and manual therapy —but is differentiated in its focus on improving a patient's ability to perform activities of daily living like bathing and dressing. SLP aims to improve patient's ability to speak and swallow. Which type of therapy a patient is prescribed depends on the patient's specific goals and medical needs. While there are similarities between the fields, each has its own specific credentials and licensing standards required to practice and bill for services in that field.

[5]Medicare also pays for therapy services in the inpatient or SNF setting through Medicare Part A. These services are not billed separately, but instead bundled into prospective payments and thus less susceptible to overspending.

[6]It is also possible for a patients to be evaluated by a physical therapist before seeing a physician (also known as "direct access"); however the physical therapist still needs to obtain a physician referral or signature at some point in order to receive payment for these evaluation services, The plan of care must be recertified within 90 calendar days from the date of initial treatment, or anytime the patient's long-term goals must be revised.

Therapy Association, 2020). Regardless of the type or setting of the services, outpatient PT is subject to the typical 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.

Policymakers have been concerned about medically unnecessary overuse of therapy services in the Medicare program for decades. In 1993, the Health Care Financing Administration (now the Centers for Medicare and Medicaid Services, or CMS) launched a task force to address widespread reports of overbilling for therapy (US General Accounting Office, 1996). In the years following, many policies were put into place to try to curb therapy spending, including setting limits on quantity through caps as discussed in the next section. But by 2018, an Office of Inspector General (OIG) audit still 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 (Office of the Inspector General, 2018).

According to the OIG, the most common reason for unnecessary PT services was due to patients receiving excessive amounts of therapy – that is, for patients who may have initially needed therapy, the amount, frequency, or duration they received went beyond "standards of practice."[7] Thus, efforts to curb excessive PT spending have focused on limiting the amount of per-patient spending in the form of annual "therapy caps."

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. 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.[8] The 1999 cap

[7] 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 (Centers for Medicare and Medicaid, 2020)

[8] 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)

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,[9] but the cap was highly salient to providers. As a result of aggressive lobbying by the PT industry,[10]. 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). Congress placed a 2-year moratorium on the cap in 2000, and in subsequent years continued to included provisions to extend the delay of the cap one year at a time.

After 6 years of Congressional extensions, the therapy cap was reinstated as two $1740 caps in January 2006, with a major modification. Medicare introduced a process through which providers could request exceptions for medically necessary services above the cap, thus making it a "soft" cap. When billing for services that would push a patient above the cap, they were supposed to indicate that they had documentation justifying a patient's medical necessity by using a billing code called the "KX" modifier code. Providers did not have to attach the documentation to their claim; 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% of claims in which the modifier documentation should have been used but was not were still paid out in full in 2006.

Medicare required that the following types of medical necessity documentation be available upon request: 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 potential separate justifications to indicating the reasoning for services that are "more extensive than is typical for the condition treated" (Centers for Medicare and Medicaid, 2024). 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 dura-

[9]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.

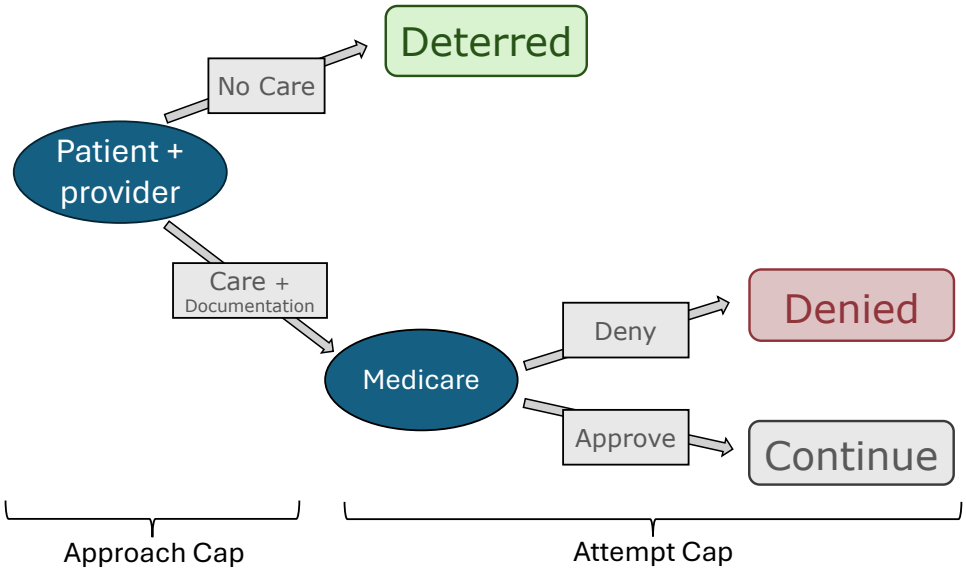
[10]Physical therapy industry representatives reportedly launched "feverish lobbying campaigns ... directed at softening" the payment changes brought on by the Balanced Budget Act (Goldstein, 1999)

tion for the particular needs of the patient, as well as relevant information on the patient's functional abilities.

The cap was enforced through Medicare claim denials, as shown by the increase in denial rates at the cap that appears in 2006 in Figure B2. In order for the therapist to charge the patient in the event of a 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 item 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 (Centers for Medicare and Medicaid, 2013, 2018).

Figure 1 shows the actions and outcomes related to the therapy cap for patients, providers, and Medicare. We will introduce several terms that will be used throughout the rest of the paper at a high-level here, and describe how they are identified in the data in Section 2.3. We consider provider behavior at two key inflection points in a patient's care trajectory: the week in which they "attempt" to cross the cap 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 but without an attempt are considered "deterred" by the cap. Alternatively, they can choose to continue care and attempt to bill above the cap. Medicare then decides whether to approve or deny this attempt. Patients who stop at the cap after attempting are considered "denied." In total, patients stopped by the cap comprise those who approach and were deterred and those who attempted and were denied. This distinction between deterrence and denial is important because they capture two different channels through which the cap curbs spending. Through deterrence, providers and patients decide whether to try to continue care after assessing their private costs and benefits. In contrast, through denials, Medicare plays an active role in deciding who is allowed past the cap. We will seek to empirically measure the size of each respective channel and also characterize their targeting properties.

Figure 1: Diagram of Therapy Cap Actions and Outcomes



Footnote: This figure illustrates patient, provider, and Medicare's actions and outcomes at the therapy cap as discussed in Section 2.2.

2.3 Data and Sample

Data Our main source of data is the Medicare 20% Carrier claims files in 1999-2000 and 2004-2008. These data include all Part B office-based spending for a random 20% subset of traditional fee-for-service Medicare beneficiaries. At the line-level, these data include procedure (HCPCS) codes, units, final payments, and indicators for denial. At the claim-level, these data include diagnosis codes (ICD-9), billing modifier codes, dates of service, and provider identifiers. We define a "provider" as a unique combination of tax identification number (TIN) and state.[11] We use TINs to define providers as opposed to National Provider Identifier (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 investments or practices that providers would take up in response to the cap would be implemented at a

[11] 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.

practice-level, as opposed to at the individual provider level. These include, for example, any upgrades electronic medical record systems, improvements to documentation standards, or changes in screening practices. In 2008, the average TIN is our sample is associated with 2.5 NPIs.

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 files.

Sample Definition We then narrow down to a sample of patients who receive in-office physical therapy[12] We follow Amico et al. (2015) in identifying PT claims via the HCPCS code and by the PT modifier code. We then focus on the 91% of these patients who only see one provider for PT the entire year. Among these patients, we further restrict to patients with "regular" PT: those that 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 (i.e, 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.

Table 1 reports summary statistics for the different samples used in each set of analyses 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 sample in Section 4. Overall, the 2005 and 2006 samples are relatively balanced on measures of demographics, prior utilization, and predicted spending.

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

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 ($)	67699	68735	67166	68131	70668	71508	71216	71910
Low-income subsidy	0.15	0.15	0.15	0.14	0.17	0.18	0.17	0.19
Dual eligible	0.14	0.15	0.14	0.22	0.16	0.17	0.16	0.18
Utilization and Spending								
Medicare payment past year ($)	9156	10014	10384	10067	11224	12530	11606	13228
Any hospital stay last 6m	0.22	0.23	0.24	0.22	0.26	0.27	0.27	0.28
In-office spending last 6m ($)	2088	2249	2313	2320	2538	2771	2620	2912
PT-related surgery last 6m	0.10	0.17	0.20	0.22	0.11	0.21	0.12	0.21
Pain procedure last 6m	0.31	0.33	0.35	0.36	0.32	0.34	0.32	0.34
Predicted 12-month PT spending ($)	.	.	1450	1473	1618	1703	1649	1760
PT utilization/spending								
Number of visits	12.7	12.7	11.7	10.7	23.0	20.6	25.2	23.7
Number of weeks of PT	7.7	7.8	7.1	6.6	13.4	12.4	14.6	14.2
Total PT spending ($)	1161	1114	1035	890	2426	2030	2688	2345
Average spending per visit ($)	105	103	99	96	127	120	129	119
Week of year start PT	22	21	26	26	14	14	14	14
Number of beneficiaries	125142	128841	62063	71019	38209	35923	31082	25126
Number of providers	13222	13084	9242	9520	8928	8784	8164	7354

This table reports summary statistics for four 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 sample in Section 4. Overall, the 2005 and 2006 samples are relatively balanced on measures of demographics, prior utilization, and predicted spending. Data: 20% Medicare Carrier claims, Master Beneficiary files, MEDPAR, and Outpatient files.

Denials Each claim is comprised of several line items, where each line is associated with a procedure code and an indicator for denial. We classify a line item as denied if the Carrier

claim payment denial code indicates a denial.[13] 94 percent of denied lines for PT in 2006 are associated with payment, while many of the remaining lines likely consist of partial payments.

Given the prevalence of denials in our data, we distinguish between two ways to summarize the spending associated with a claim: the *billed* and the *paid* amount. The amount paid reflects the payment the provider receives, 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 payment (on claims without a denial) for that procedure code.[14]

Documentation Measure While we cannot directly observe documentation in the claims data, we proxy for it using a billing modifier code (the "KX modifier" code) which providers use to indicate that they attest to the medical necessity. This modifier code was introduced as part of the 2006 therapy cap, and is used by providers to indicate the availability of documentation to justify the medical necessity of spending over the therapy cap. Documentation plays a key role whether a patient goes past the cap. As shown in Table 5, having documentation, as proxied by the presence of the documentation modifier code on the claim, lowers the likelihood that a patient is stopped by the cap, conditional on their provider making an attempt. This correlation is robust to the inclusion of controls for patient demographics and health status. The R^2 increases 3-4-fold once an indicator for having documentation is included in the regression.

Weeks from Cap We aggregate claims up to the week level and conduct much of our analyses at the patient-week level. We only consider weeks in which a patient received PT

[13] See ResDac (last accessed 3/8/24) for more detail on this code.

[14] 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).

care, and in each week we assess how many weeks of care away a patient is from reaching the cap. The decision to aggregate to the week level is motivated by patterns in the data which suggest that care is scheduled on a weekly basis. Specifically, Figure B1 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.

The method for calculating how many "weeks from the cap" a patient is differs depending on whether a patient's cumulative spending at the end of the year fell above or below the cap. First, for patients who end the year above the cap, week 0 is defined as the week of their attempt, and all other weeks are defined relative to that week.[15] Second, for patients make an attempt but never get past 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 to cross the cap (and thus end the year below the cap), we extrapolate to calculate how many "weeks away" the patient is from reaching the cap. We do this by using the patient's spending in prior weeks. Specifically, calculate the patient's 5-week rolling average of PT spending. We then 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. 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.

When "weeks from cap" is negative, it denotes how many additional weeks of care a patient would have to receive to reach (and pass) the cap. A claim submitted in the week that an unsuccessful attempt was made is considered to be -1 weeks from the cap. When "weeks from cap" is positive, it denotes how many weeks ago the patient passed the cap. Figure B2 illustrates week-level denial rates for claims depending on how many weeks away they are from the cap and shows a clear increase in the share of weeks of care with a denied claim.

[15] If a patient ends up past the cap but only after making multiple attempts, we consider their *first* attempt to be week 0.

Attempts vs. Approaches To define weeks with attempts, we first identify weeks in which a patient's cumulative *paid* amount is below the cap at the beginning of the week. Among these, a week is considered an attempt if their cumulative *paid* or *billed* amount is above the cap at the end of the week. Approaches are defined in two ways, depending on whether a patient ever attempts. First, for patients who have an attempt at some point that year, we define an approach as the last week before their attempt. If patients have multiple attempts, we use the last week before their first attempt. Second, for patients who never make an attempt that year, we extrapolate how many "weeks from the cap" they are in that week using their prior spending. We discuss this calculation in further detail below. Weeks in which the patient is estimated to be 1 week away from the cap (i.e., weeks from cap is -1) are considered approaches as well.

Deterred, Denied, and Approved In Sections 3.1 and 4, we consider patients who attempt or approach the cap, and distinguish between two reasons why patients stop at the cap. We consider a patient to be "deterred" if they approach the cap but never make an attempt. We consider a patient to be "denied" if they make an attempt but never successfully make it over the cap by the end of the year – that is, their total billed amount at the end of the year is less than the cap.[16] In Section 5, we switch from a patient-level analysis to an *attempt*-level analysis. A attempt is considered "approved" if the patient receives no denials on any PT claim in the week of an attempt. Patients classified as "denied" at the end of the year must have had at least one attempt with a denial, but not all patients with a denial in an attempt end the year denied. This is because because providers may make multiple attempts with the same patient if a prior attempt is denied.

Predicted PT Spending We also use the claims data to construct a measure of *ex ante* patients' PT need. We predict a patient's 12-month PT spending given their spending and utilization patterns prior to their first PT, which we then use to proxy for patient need in Sections 3 and 4. We use gradient-boosted decision trees from the LightGBM package. The predictors are: age, sex, utilization and spending in the previous calendar year

[16] See further discussion on the categorization of deterred and denied claims in Appendix Section C.

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.[17] The model is trained on patients who approach or attempt 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 to patients who approach and attempt the cap in 2005 and 2006. Interpreting predicted pre-reform spending as a measure of underlying need requires assuming that a patient's relative position in the spending distribution is informative about their true need, and that this ranking is stable between 2004-2005 and 2006. We discuss the machine learning methodology and model fit in further detail in Appendix Section E.

3 Overall Effects of the Cap

In this section, we will first quantify the direct effect of the therapy cap on overall Medicare PT spending. To place these savings into context, we will compare the implied savings from the 2006 "soft" cap to a one-time "hard" cap implemented in 1999. We will then consider the overall health effects of these cap-induced spending reductions in a research design that leverages the fact that the cap resets annually in January.

3.1 Savings

Methodology To quantify the Medicare savings from the cap, we will use methods from the "bunching" literature (Kleven, 2016). In contrast to the standard bunching setup, where the counterfactual distribution in the absence of the policy has to be constructed, in our context we can directly use the pre-policy distribution as our counterfactual. In particular,

[17] 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).

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 – known in the bunching literature as the "manipulation region." Medicare pays for 80 percent of allowed charges and patients are responsible for the remaining 20 percent, so the cap appears at $0.8 \times \$1740 = \$1392$ in the distribution of per-patient Medicare PT spending. We then restrict to a region from \$800 below below to \$1600 over this amount. Under the identification assumptions described below, the difference in spending between the two distributions can be interpreted as the savings implied by the cap.

Rather than directly comparing the 2005 and 2006 distributions, we adjust the 2005 spending calculation to make it a more realistic counterfactual by accounting for changes in procedure prices and in the total number of patients receiving PT over time. 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, and p_k is a vector of prices for each procedure in year k. 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.

To calculate the savings from the cap, we construct $\bar{r}(F_{05}, p_{06})$ by replacing the 2005 procedure prices with the average price the provider received for that procedure in 2006.[18] 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. This ensures that the savings calculation is not affected by secular increases in the total number of patients receiving PT in this period. Thus, the savings in 2006 dollars is defined as:

[18] We use the *provider*-level 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, 2022)

$$S_{06} := N_{06} \left(\bar{r}(F_{05}, p_{06}) - \bar{r}(F_{06}, p_{06}) \right),$$

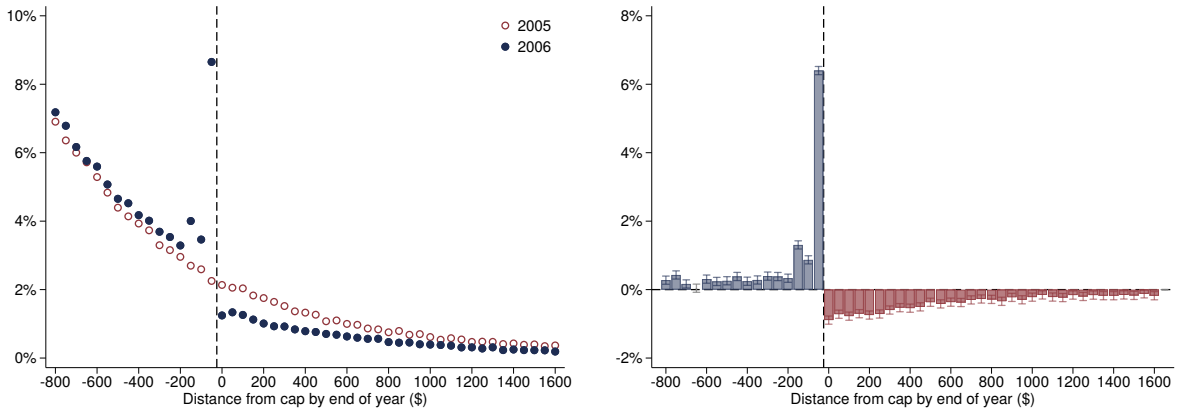
and the savings as a percent of 2005 spending is defined as:

$$S^{\%}_{06} := \frac{\bar{r}(F_{05}, p_{06}) - \bar{r}(F_{06}, p_{06})}{\bar{r}(F_{05}, p_{06})}. \tag{1}$$

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. The savings to Medicare per additional denial due to the cap was $156.

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

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



This figure plots the (a) 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 adjusted 2005 distributions as the savings from the cap requires making two 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 even without changes in Medicare policy. We can validate this by comparing the 2005 distribution to the 2004 distribution which also did not have a therapy cap. Appendix 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 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, we need that there was no "extensive margin" response to the cap which differentially drew patients into or out of the manipulation region. This assumption allows us to normalize patient count across the two years by just multiplying by the 2006 patient count. 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 with relatively minor injuries who would be far below the cap. In this case, the relative share of patients outside of the manipulation region would increase once the reform is in place.

We substantiate this assumption in two ways. First, looking at the full distribution of spending in 2005 and 2006 in Appendix 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. Furthermore, we would expect that providers with relatively more patients over the cap should would face the largest capacity expansions as a result of the cap; thus, any increases in low-spend patients should be concentrated among these providers. Appendix Figure B4 panel (b) plots the average patient count, splitting by providers who had high vs. low shares of 2005 patients over the cap. Providers with high 2005 over cap shares do not appear to see more patients start in 2006, indicating that they did not respond to the cap-induced capacity expansion by taking on more low-spend patients.

Comparison to a Hard Cap To put the savings from the 2006 soft cap into context, we 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.[19] Due to data limitations, we

[19] 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).

cannot compare 1999 to a pre-period year but instead compare it to 2000, after the cap was repealed. Any lingering effects of the 1999 cap in 2000 spending, 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 in 2006, this confirms that the 2006 cap was indeed "soft" – the savings are diminished by almost half when providers are allowed to request exceptions to exceed the cap.

Distinguishing Deterrence and Denial Channels Figure 2 shows that the therapy causes patient spending to bunch at the cap, indicating that patients are stopping just short of the cap. To unpack what is driving this bunching, we next turn to examining how providers' decisions to continue care and Medicare's decisions to deny claims change as they approach the cap.

In order to accurately capture when patients "stop short" of going over the cap, we ask how many *weeks* of the patient's "typical" care it would take for them to reach the cap. We use weeks instead of dollars for two reasons. First, it may be difficult for providers to target an exact dollar amount, since patients receive discrete units of care with fixed prices and also vary in the type of care that they need. Second, the week-to-week correlations in day of care in Figure B1 indicate that therapy visits are scheduled on recurring days of each week. Furthermore, they also exhibit a Monday-Wednesday-Friday and Tuesday-Thursday scheduling pattern. This suggests that the decision to continue care is made on a week-to-week basis.

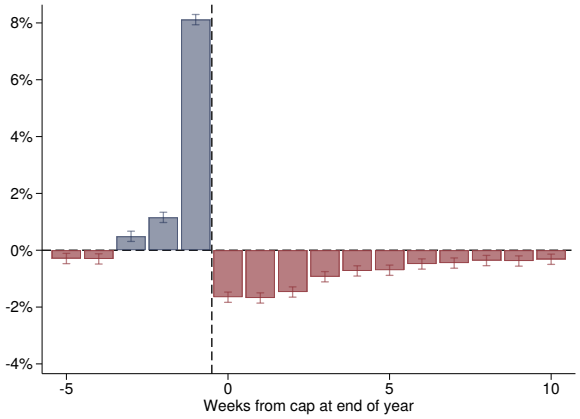
For each patient-week observation we calculate how many weeks away they are from the

cap. The week in which a patient crosses the cap, defined as the week their *cumulative* reimbursement surpasses the cap, is assigned week 0. Positive weeks are past the cap, and negative weeks are prior to the cap. For patients who pass the cap, positive weeks are defined relative to week 0. For patients who never cross the cap, we extrapolate how many weeks it would take to cross the cap based on their last 5 weeks of spending. Section D describes the definition of weeks to cap in further detail.

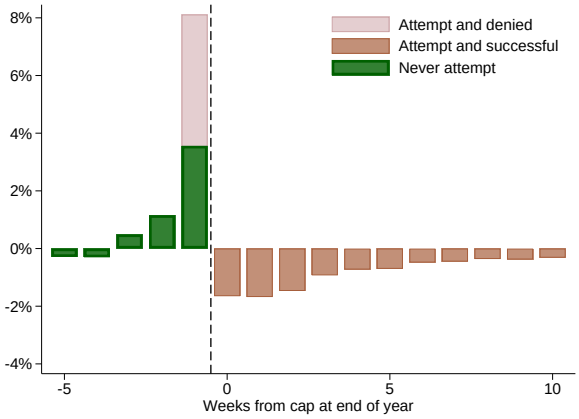
In Figure 3, we plot the distribution of end-of-year PT spending by the number of weeks a patient is from the cap. Panel (a) plots the difference between the 2005 and 2006 distributions. The excess mass to the left of the cap is concentrated in the last 3 weeks leading up to the cap, and there is a 93 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, with effects extending even up to 10 weeks away from the cap.

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

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



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



This figure plots the difference in the 2005 and 2006 distributions of cumulative physical therapy. 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). The patients below the cap who "never attempt" represent those who are stopped due to deterrence channel, while those who categorized as "attempt and denied" represent those who are stopped due to the denial channel. Section 2.3 describes the construction of the weeks from cap measure as well as the definitions of deterrence and denial in further detail. Data: 20% Medicare Carrier claims.

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 by Medicare and thus end the year one week below the cap, or they are classified as "attempt and successful," meaning they made an attempt that was not denied by Medicare and thus end the year above the cap. The remaining patients are classified as "never attempt," and end the year one or more weeks below the cap. The deterrence channel is captured by patients to the left of the cap who never attempt, and the denial channel is captured by those who attempt and are denied. The patients below the cap who "never attempt" represent those who are stopped due to deterrence channel, while those who categorized as "attempt and denied" represent those who are stopped due to the denial channel. Of patients who bunch in the week *immediately*

before the cap, 43 percent stop due to deterrence while 56 percent stop because they are denied. While most of the excess mass appears exactly one week short of the cap, there is also a small amount of excess mass dispersed up to three weeks away from the cap. Once patients who stop up to three weeks away are included, 53 percent stop due to deterrence while 47 percent stop due to a denial.

Comparing the distribution of beneficiaries allows us to estimate the share of *beneficiaries* who stop at the cap into the denial and deterrence channels. However, the decomposition of *savings* due to each channel could differ, as it depends not only on the number of beneficiaries stopped but also these patients' 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. We use the predicted 2004-2005 PT spending measure described in Section 2.3 and Appendix Section E.

Figure B6 plots the distributions of predicted PT spending among "deterred" and "denied" patients who stop one week before the cap in 2005 (panel a) and 2006 (panel b), as well as the difference between the two distributions (panel c). The savings through each channel is represented by the sum of the area under the 2006 distribution minus the area under the 2005 distribution. 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.[20] This split is fairly close to the share of beneficiaries who are deterred vs. denied, reflecting the fact that the medians and spread of the predicted savings distributions are fairly similar between the deterred and denied groups.

3.2 Patient Health Effects

We next consider the overall efficacy of the cap by looking at whether it had a direct effect on patient health or led patients to substitute to other forms of care. If there is evidence of direct patient harm or substitution to alternative (and possibly more costly) types of care, this suggests that the cap-induced spending reductions were not well-targeted. To uncover the causal effect of the therapy cap on health outcomes, we use an instrumental variables

[20] If beneficiaries who end the year two and three weeks away from the cap are included, then the savings from deterrence rise to $54 million, or 51% of the total savings.

(IV) strategy that leverages within-year variation in spending due to the differential "bite" of the cap depending on when a patient initiates PT. Appendix Section G describes the methodology and results for an alternative identification strategy which employs a difference-in-difference research design; this alternative design finds similar results as the IV design.

Methodology Our identification strategy relies on the fact that the therapy cap resets at the beginning of each calendar year, regardless of when they began receiving PT care in the prior year. All else equal, the cap should bind more for a patient who starts their care earlier vs. later in a year. We would expect that the relationship when a patient starts PT in the year and their total 12-month PT spending to become more negative once the cap is in place. This motivates our instrument: the month that a patient starts PT, interacted with year.

We construct our sample by first taking patients who start PT in either 2005 or 2006, where "starting" PT session is defined as having no PT in the six months prior. In order for the timing of when patients start PT to be a valid instrument for PT spending, we need to assume that exclusion restriction holds: the only reason for the relationship between start month and a patient's health is because patients who start earlier in the year are more likely to have their PT spending constrained by the cap. One potential violation of this assumption would be if the cap changes the composition who starts care earlier vs. later in the year. For example, if healthier patients were more aware of the cap and strategically timed their care to start later in 2006, this would result in a negative correlation between patient health and PT start date that is not causal. To address this concern, we restrict to a patient population where the PT start date is more likely to be exogenous: those who received an injury diagnosis in the 90 days prior to their first PT visit.[21] Finally, we restrict to patients with 12-month PT spending over $200 to focus on patients more likely to be affected by the therapy cap.[22] As robustness, we re-run our analyses on patient populations who may be more vulnerable to reductions from the therapy cap: low-income patients and high-need patients. We also look for heterogeneous effects by decile of predicted patient need

[21] Appendix Section F describes how injuries are identified in the sample construction.

[22] 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.

and whether a patient had a pain procedure or orthopedic surgery prior to starting PT.

We first estimate the following reduced form specification:

$$Y_i = \sum_{m=1}^{m=11} \theta_m 1(Year_{y(i)} = 2006) \times FirstMonth_{m(i)} + \alpha_{m(i)} + \lambda_{y(i)} + \varepsilon_i, \tag{2}$$

where the omitted month is December. $Year_{y(i)}$ is the year a patient starts PT and $FirstMonth_{m(i)}$ is the month of a patient's first PT. The specification for the first stage of our IV strategy is:

$$PT_i = \gamma_0 + \gamma_1 1(Year_{y(i)} = 2006) \times FirstMonth_{m(i)} + \nu_i, \tag{3}$$

where PT_i is the patient's 12-month PT spending. Then the second stage is:

$$Y_i = \beta PT_i + \alpha_{m(i)} + \lambda_{y(i)} + \varepsilon_i, \tag{4}$$

where Y_i is a health or utilization measure, and $\alpha_{m(i)}$ and $\lambda_{y(i)}$ capture first month and year fixed effects We will scale PT_i so that β can be interpreted as the effect of an additional \$100 in Medicare PT spending. Our estimates are clustered by year and start month.

In addition to the 12-month PT spending, we consider six indicators of patient health and utilization that could be plausibly related to an insufficient amount of PT: (1) opioid prescriptions, (2) pain management procedures, (3) orthopedic surgeries, (4) emergency department visits, (5) inpatient stays, and (6) 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 requires 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 after the last day to capture patients seeking alternatives after PT has "failed," as opposed to utilization that

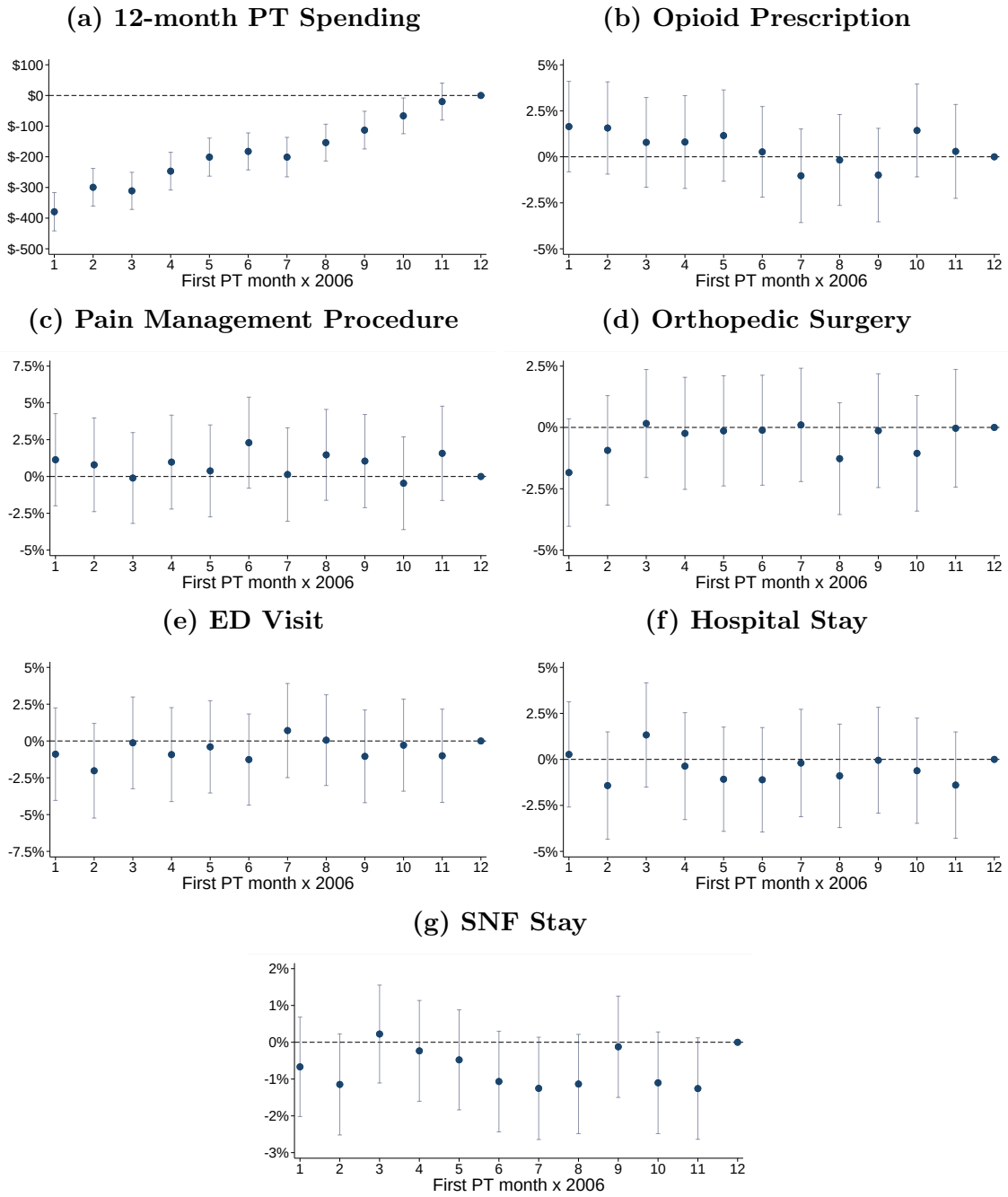
could be endogenous to the PT's behavior – say, if a PT is co-located with a pain management practice. The opioid prescription outcome is an indicator variable that is measured between 12 and 24 months after the patient's last day of PT. This is because the Part D prescription data for opioids is only available in 2006, so we cannot measure opioid prescriptions within 12 months for the 2005 sample. Appendix Section F 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-funded PT spending, so our estimates tell us causal health effects of reducing *Medicare*-funded PT spending, rather than the effects of reducing PT spending overall. If patients fully compensate for the reductions in Medicare spending by paying for the rest of their care out-of-pocket, then the therapy cap just represents a reduction in transfers, but not in PT utilization, for patients. In practice, patients likely do not fully compensate for these reductions, so the therapy cap should reduce overall PT spending. However, since our data is restricted to Medicare claims, we cannot measure how much a reduction in Medicare-funded PT spending passes through to overall spending. We can address this partially by repeating our analyses just on low-income patients, for whom we might expect more passthrough. Additionally, since we are using claims data, we cannot observe some dimensions of health that arguably PT is most relevant for, such as mobility and pain levels. Our analysis cannot speak to whether patient well-being worsens in these dimensions unless the effects are large enough to induce substitution to other forms of care or trigger events like ED visits.

Results Figure 4 shows the reduced form results from Equation 2. Panel (a) plots the relationship between first month of PT interacted with 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 mostly monotonic relationship start month and 12-month PT spending. The 2005-2006 difference in PT spending is much larger for patients who start earlier in the year relative to those who start later. Turning to our health outcome measures, we would expect that if the cap-induced spending reductions were poorly targeted, then there would a negative relationship between these outcomes and a patient's start month. Instead, this

relationship is flat and nearly all coefficients are statistically insignificant.

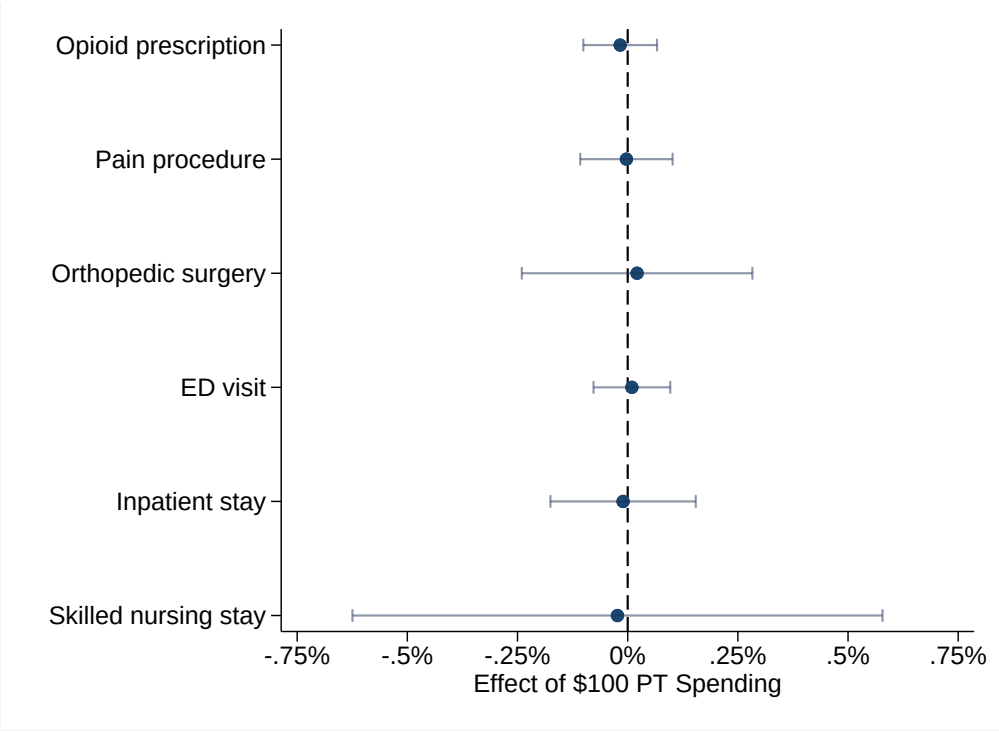
Figure 4: Reduced Form Spending and Health Outcomes



This figure plots the coefficient θ_m, which denotes the interaction between an indicator for 2006 and an indicator for month of first PT, from Equation 2. 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 SNF 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 F describes how the health outcome measures are defined. Sample is restricted to beneficiaries with less 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 Equations 3 and 4, which are scaled to represent the causal effect of an additional $100 in PT spending on each outcome, relative to the 2005-2006 average. Given our relatively strong first stage,[23] we estimate precise null health effects. We are able to rule out out effect sizes larger than 0.25 percent for opioids, pain procedures, and orthopedic surgeries, effect sizes larger than 0.5 percent for orthopedic surgeries, and effect sizes larger than 0.75% for skilled nursing stays.

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



This figure plots the coefficient β, which denotes the effects of an additional $100 of PT, from Equation 4. 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 and are scaled in percent terms relative to the sample average. Section F describes how the health outcome measures are defined. Sample is restricted to beneficiaries with less than $200 in 12-month PT spending. Data: 20% Medicare Carrier, Outpatient, MEDPAR, and Part D claims.

The lack of health impacts on the overall sample could be masking effects for particularly vulnerable patients. We address this in three ways. First, we also focus on two patient populations where the reduction in Medicare PT spending could be particularly harmful. We first consider low-income patients, who may be unwilling or unable to pay for additional

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

PT care out-of-pocket in the face a Medicare denial. If so, the reductions in Medicare PT spending from the cap should translate into relatively larger reductions in *overall* spending for this population, 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 LIS – we do not find evidence that the therapy cap worsened their health outcomes.

Next, we stratify by a patient's initial health status, which is defined as their decile of predicted 2004-2005 PT spending constructed in Section 2.3. We estimate the IV in Equation 4 separately for each decile, where higher deciles indicate a patient is predicted to have more PT spending in 2005. 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.

Finally, we stratify by whether the care the patient recently received prior to starting PT. Some patients may start PT to aid in the recovery from a more invasive treatment, such as a pain management procedure or orthopedic surgery. 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 in the 6 months prior to starting PT. 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

Thus far, we have measured the extent to which the 2006 therapy cap affected PT spending and provider behavior at the cap. The spending reductions did not worsen patient health, suggesting that they were well-targeted at low-value care. Another way to assess the effectiveness of the cap is by assessing *who* the cap screens out. The first set of analyses complements the health analysis in Section 3.2 and evaluates whether the cap targeted low-value spending. Given that its stated goal was to restrict medically unnecessary care, we compare patients of different levels of clinical need to determine if lower-need patients are

more likely to be screened out than higher-need ones. If so, this would indicate that the cap improved the efficiency of Medicare spending.

The second set of analyses examines *horizontal inequity* – that is, whether patients with the *same* level of clinical need face different probabilities of being screened out by the cap. Because the cap introduced a documentation requirement, we pay particular attention to disparities across providers with varying administrative capacity. Since administrative capacity is often correlated with provider size, we compare the effects of the cap on patients of large vs. small providers. Furthermore, since provider size is also correlated with patient demographics, we further explore whether differences in the effects of the cap across providers translated into disparities across patient groups.

Figure 3 shows that the cap operates through two distinct channels: by deterring attempts to bypass the cap or, conditional on an attempt, by denying them. The factors influencing whether a patient makes it past each stage differ – the deterrence decision reflects a joint decision to between the patient and their provider, while a denial depends on Medicare's evaluation of the documentation and enforcement of its medical necessity standard. We will thus consider the screening properties of the deterrence and the denial stages separately.

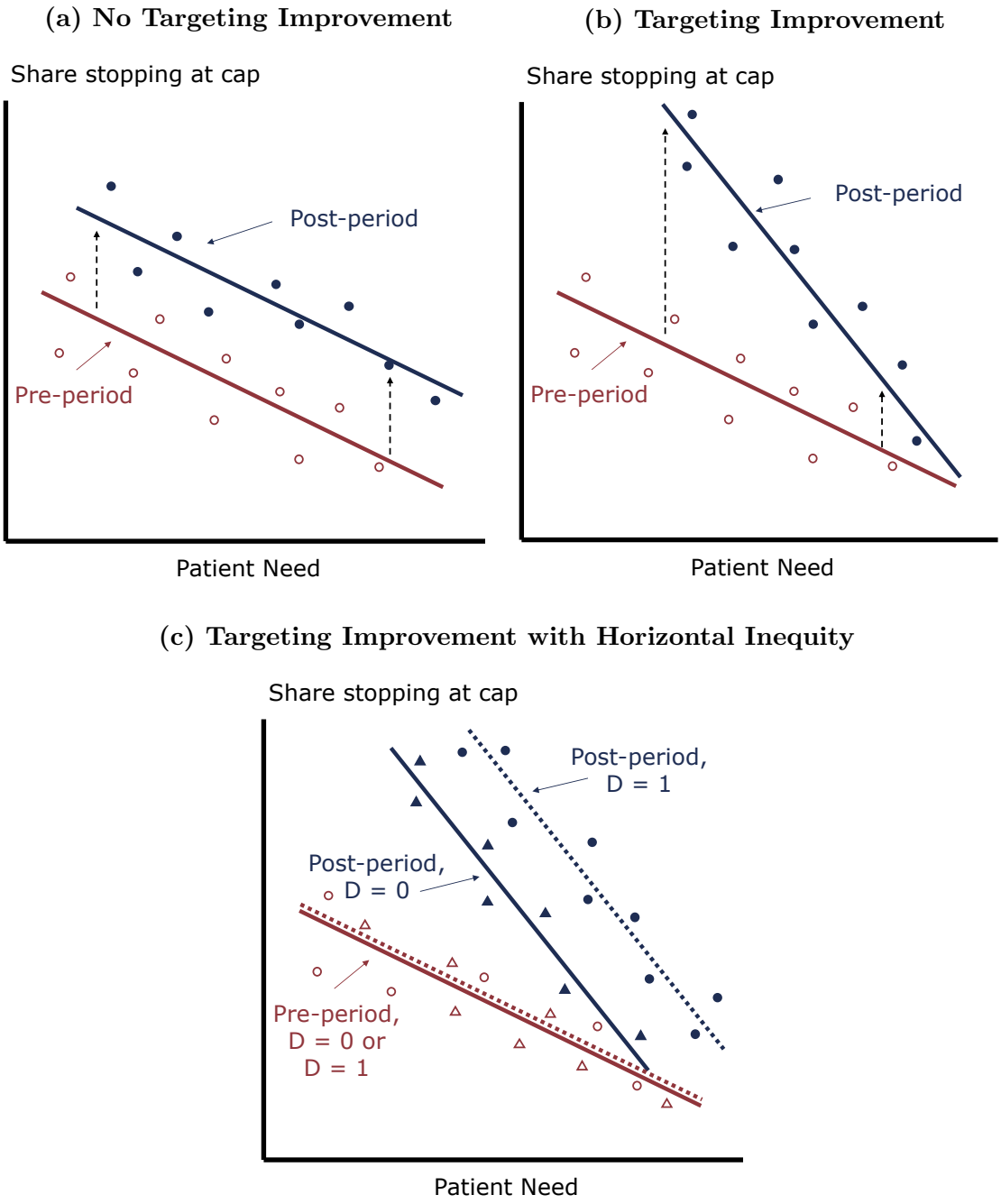
4.1 Methodology

To characterize the screening properties of the cap, we will examine the relationship between a given characteristic and the outcome of interest at each stage. We will isolate the targeting properties of the cap by looking at whether the slope of these relationships *change* between 2005 to 2006. In particular, if a correlation becomes stronger once the cap is implemented, this would indicate that the cap increased targeting on that characteristic.

Figure 6 illustrates the intuition of this exercise. In this example, patient need is plotted against the likelihood that a patient stops at the cap. Panel (a) shows the case where the cap does not improve screening, while panel (b) shows the case where it does. In 2005, the therapy cap is not in effect yet, so the "share stopping at cap" is simply the natural rate at which patients stop where the cap will eventually be located. In both panels, the downward-sloping relationship indicates that higher-need patients are less likely to stop at the cap. This relationship can be negative even before the cap is in place – for example, this

would be the case if providers give more care to sicker patients, making them more likely to stop at any arbitrary spending level for healthier ones.

Figure 6: Illustration of Targeting and Inequity Exercise



This figure presents illustrations of the screening exercise in Figures 7 and 8 . Panel (a) illustrates the case when the introduction of the cap increases the share of patients who stop at the cap (either due to deterrence or denials), but does not improve targeting. The increase in the share stopping at the cap is uniform across all levels of patient need. Panel (b) illustrates the case when the cap increases the share who stop at the cap and improves targeting. In this case, the increase in the share stopping at the cap is larger for low-need patients and smaller for high-need patients. The overall likelihood of receiving care past the cap has become correlated with need once the cap is introduced, indicating that the cap screens on need. Panel (c) illustrates the case when the cap improves targeting and also introduces horizontal inequity on characteristic D. Holding fixed patient need, the increase in share stopping at the cap is larger when $D = 1$ than when $D = 0$.

Figure 6 panel (a) illustrates the case where the cap increases the share of patients who stop at the cap but does not improve targeting on need. To see this, note that the likelihood of stopping at the cap increases uniformly at all levels of need but the slope of this relationship remains unchanged. So while all patients were more likely to stop, high vs. low-need patients were not differentially affected. In contrast, in panel (b) all patients are more likely to stop at the cap, but the increase is *largest* for low-need patients. This forms the basis for our test for increased targeting: whether the slope of this relationship strengthens between 2005 and 2006.

Figure 6 panel (c) then illustrates the intuition of our test for horizontal inequity on an indicator for dimension D (e.g., D could be an indicator for going to a large or small provider). Holding fixed patient need, patients with $D = 1$ in the pre-period, represented with the circles and dotted lines, are no more likely to stop at the cap than patients with $D = 0$. However, once the cap is in place, a gap emerges between patients with $D = 1$ and $D = 0$ – at the same level of need, $D = 1$ patients are more likely to stop at the cap. In other words, the cap has introduced horizontal inequity along dimension D.

We first present our results as binscatters (residualized of week-of-year effects) of the relationships between two separate outcomes at the cap – deterrence and denials – and a measure of patient need. When looking at deterrence, we focus on the sample of patients in the week they approach the cap, defined as Section 2.3. These are patients for whom one additional week of their regular care would have taken them over the cap. For denials, we focus on patients in the week of their attempt to go over the cap. We formally test for slope changes between 2005 and 2006 by estimating following equation:

$$Y_i = \beta_1 X_i + \beta_2 X_i \times 1(Year_{t(i)} = 2006) + Week_{t(i)} + \varepsilon_i, \tag{5}$$

where Y_i is a dummy variable either the patient was deterred or denied, X_i is a measure of patient need, and $Week_{t(i)}$ is a week-year that the patient approached or made an attempt. β_1 is the relationship between the dimension of screening we're interested in and outcome Y_i in 2005, and β_2 captures how that relationship changes in 2006. Thus, our test of whether screening increases after the cap is introduced is whether β_2 is statistically significant (and

going in the same direction as β_1).

Then to test for horizontal inequity conditional on need, we estimate the following equation:

$$Y_i = \beta_1 X_i + \beta_2 X_i \times 1(Year_{t(i)} = 2006) + \beta_3 D_i + \beta_4 D_i \times 1(Year_{t(i)} = 2006) + Week_{t(i)} + \xi_i, \quad (6)$$

where D_i is an indicator variable for whether a patient goes to a small provider or a patient demographic characteristic. After controlling for patient need, β_3 captures whether patients with $D_i = 1$ are at baseline more or less likely to stop at the cap, and β_4 captures whether this relationship changes once the cap is in place.

We use the predicted 2004-2005 PT spending measure discussed in Section 2.3 as our measure of patient need, X_i. Interpreting predicted pre-reform spending as a measure of underlying need requires assuming that a patient's relative position in the spending distribution is informative about their true need, and that this ranking is stable between 2004-2005 and 2006. In other words, even if the overall *level* of spending in 2004 or 2005 may have been considered too high by Medicare, relatively high-spend patients in 2005 have more PT need than relatively low-spend ones.

One potential limitation of using a predicted measure trained on healthcare claims is that these 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 is predictive of greater spending in the future, then the model would underestimate need for these populations. This is an issue inherent in any analysis using claims-based measures to proxy for patient need. For the first set of analyses looking at the change in slope, if predicted and true need were weakly positively correlated or completely uncorrelated with each other, then this would bias our test against finding an improvement in targeting. Under- or over-estimating true need would only undermine our analysis if the bias is large enough such that the predicted and true need were sufficiently *negatively* correlated with each other – in that case, what we would classify as an improvement in targeting (i.e., a steepening of the slope) would actually be indicative of

worsened targeting. This would be a fairly extreme case of model misspecification. For the analyses looking at horizontal inequity, this issue would generate a gap in outcomes between $D_i = 1$ patients and $D_i = 0$ patients in both the pre-reform and post-reform period. Assuming the magnitude of this error does not change over time, we would not expect it to affect our coefficient of interest β_4, which is on the interaction between D_i and an indicator for $Year_{t(i)} = 2006$. Finally, we further bolster our results by repeating our analysis using age in Appendix Figure B12, which is plausibly correlated with true need and less prone to claims-based measurement error.

4.2 Results: Targeting on Need

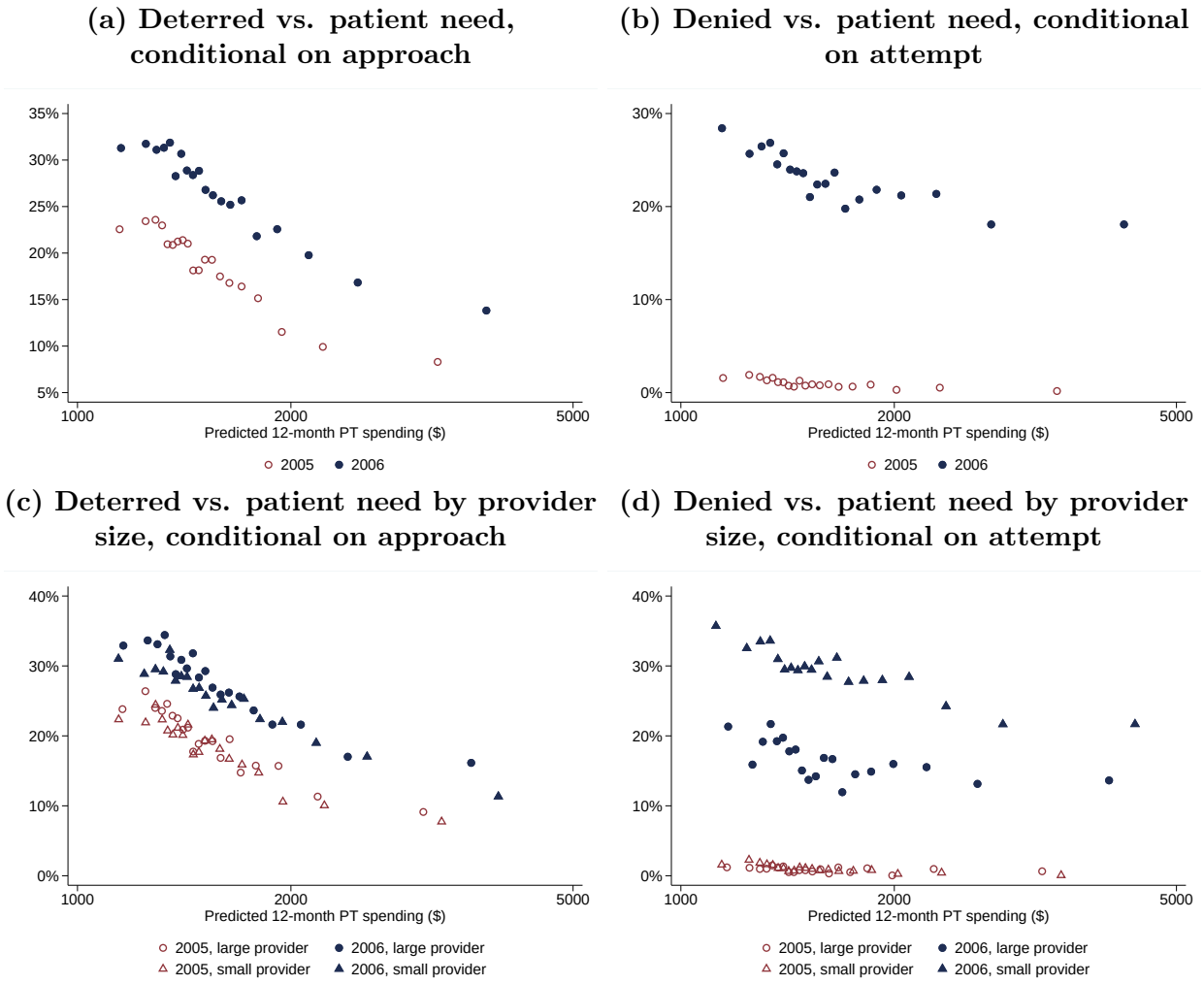
Figure 7 panel (a) plots the correlation between patient need and deterrence. Both before and after the cap, PTs are less likely to be deterred at the cap for patients with higher need. Once the cap is in place in 2006, they are noticeably less likely to continue to attempt care for patients – on average, there is 64 percent (11.9 percentage point) increase in the likelihood that a patient stops at the cap without an attempt. However, the magnitude of this increase is similar across all levels of need – this is reflected in the uniform upward shift in levels and the lack of a change in its slope. In the aggregate, it does not appear that the cap leads to more screening on need in the deterrence stage, where patients and providers are deciding whether to attempt to go past the cap.

Panel (b) then considers how likely a patient is to be denied by Medicare, conditional on making an attempt to bill over the cap. In contrast to the deterrence results in panel (a), here we see that the cap leads Medicare to be much more discerning in the care it approves. Prior to the cap, Medicare rarely denied attempts, and the denial rate only has a slight negative correlation with need. Once the cap is introduced, Medicare became noticeably more likely to deny care for all patients, as shown in the shift up in denial rates at all levels of need. The magnitude of this increase differs by patient need – it is substantially larger for low-need patients, as demonstrated by the steepening of the slope.

We confirm this formally in Table 2, which shows that the coefficient on "Predicted spending $\times$ 2006" is insignificant for deterrence (column 1) but statistically significant and negative for denials conditional on attempt (column 3). Column 5 shows that net effect on

whether a patient eventually gets past the cap, conditional on *approach*, is an improvement in targeting, as the coefficient for denial conditional on approach on "Predicted spending × 2006" is negative.

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



This figure plots the relationship between attempt outcomes and patient need and provider size. "Deterred" is defined as the share of patients who approach the cap but do not attempt, and "denied" is defined as the share of patients who attempt but never make it past the cap. Panels (a) and (b) plot the relationship between log predicted 12-month PT spending and share deterred and denied 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 2.3 and Appendix Section E discuss the construction of the predicted PT spending measure and Section 2 describes the sample definition in further detail. Data: 20% Medicare Carrier claims and Master Beneficiary Summary File.

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

	(1)	(2)	(3)	(4)	(5)	(6)
	Deterred		Denied			
	Approach		Attempt		Approach	
Predicted Spending	-17.3***	-17.3***	-1.94***	-1.95***	-19***	-18.9***
	(.71)	(.712)	(.219)	(.22)	(.738)	(.739)
Predicted Spending × 2006	-3.44***	-3.45***	-9.63***	-10.1***	-7.25***	-7.56***
	(1.09)	(1.09)	(.923)	(.92)	(1.25)	(1.25)
Small provider		-1.11**		.3*		-.869
		(.526)		(.176)		(.551)
Small provider × 2006		.874		10.3***		7.77***
		(.888)		(.844)		(1.08)
Outcome mean, 2005	18.7	18.7	1.4	1.4	19.9	19.9
Outcome mean, 2006	30.6	30.6	20.5	20.5	44.4	44.4
Week-year FE	X	X	X	X	X	X
Cluster	TIN	TIN	TIN	TIN	TIN	TIN
N. Clusters	11889	11889	11057	11057	11889	11889
Observations	74024	74024	56111	56111	74024	74024

This table presents the coefficients from estimating Equation 5 with log predicted PT spending as X_i (columns 1, 3, and 5) and Equation 6 with log predicted PT spending as X_i and an indicator for whether a patient goes to an above-median-size provider as D_i (columns 2, 4, and 6). The coefficient β_1 is the estimate for "Predicted spending," the coefficient for β_2 is the estimate for "Predicted spending × 2006", the coefficient for β_3 is the estimate for "Small provider" and β_4 is the estimate for "Small provider × 2006." The regression is clustered at the provider (TIN-state) level. Provider size is measured as a TIN-state's 2006-2008 Medicare patient count and "small provider" denotes a provider with below-median patient count. Columns 1-2 and 5-6 restrict to the sample of 2005-2006 patients who ever approach the cap and columns 3-4 restrict to the sample of patients who ever make an attempt to go past the cap, as defined in Section 2.3. Section 2.3 and Appendix Section E discuss the construction of the predicted PT spending measure and Section 2 describes the sample definition in further detail. Data: 20% Medicare Carrier claims.

Taken together, these results show that while the overall savings from the cap are due to both increased provider deterrence and Medicare denials, the improvements in targeting on need come only from *Medicare*'s scrutiny. The extent to which providers pulled back their attempts after the cap was introduced did not differ by a patient's *ex ante* need, suggesting that the ordeal itself was a fairly blunt targeting tool.

Given the potential concerns with using a claims-based prediction as an index of need, as additional robustness we repeat the analysis and estimate Equation 5 using patient age as our measure of need. The results are reported in Figure B12 panels (a) and (b), and Table A2 columns (1), (3), and (5). They show patterns consistent with the main results: the cap improves screening only in the denial stage.

4.3 Results: Differences by Provider Size and Patient Demographics

By Provider Size We next consider whether the cap also screens on the type of provider a patient goes to, in addition to that patient's underlying need. We focus on differences by provider size as prior work has shown that larger providers tend to have an advantage in billing: Dunn et al. (2024) show that smaller providers incur higher billing resubmission costs when appealing claim denials, and League (2022) has shown that one way providers respond to increases in insurer denials is by consolidating. At the deterrence stage, larger providers may be better at selecting patients if there are returns to experience or to scale in the ability to assess patient need. For example, larger providers may have access to better information about a patient's medical history or have processes to more systematically assess patient need. At the denial stage, larger providers may also have a technological or scale advantage in terms of keeping documentation, thus increasing their success rate on attempts.

Figure 7 panels (c) and (d) plot the same correlations as before between patient need and deterrence or denials, 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."[24] Panel (c) shows that, at a given level of need, there is no difference in deterrence across large or

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

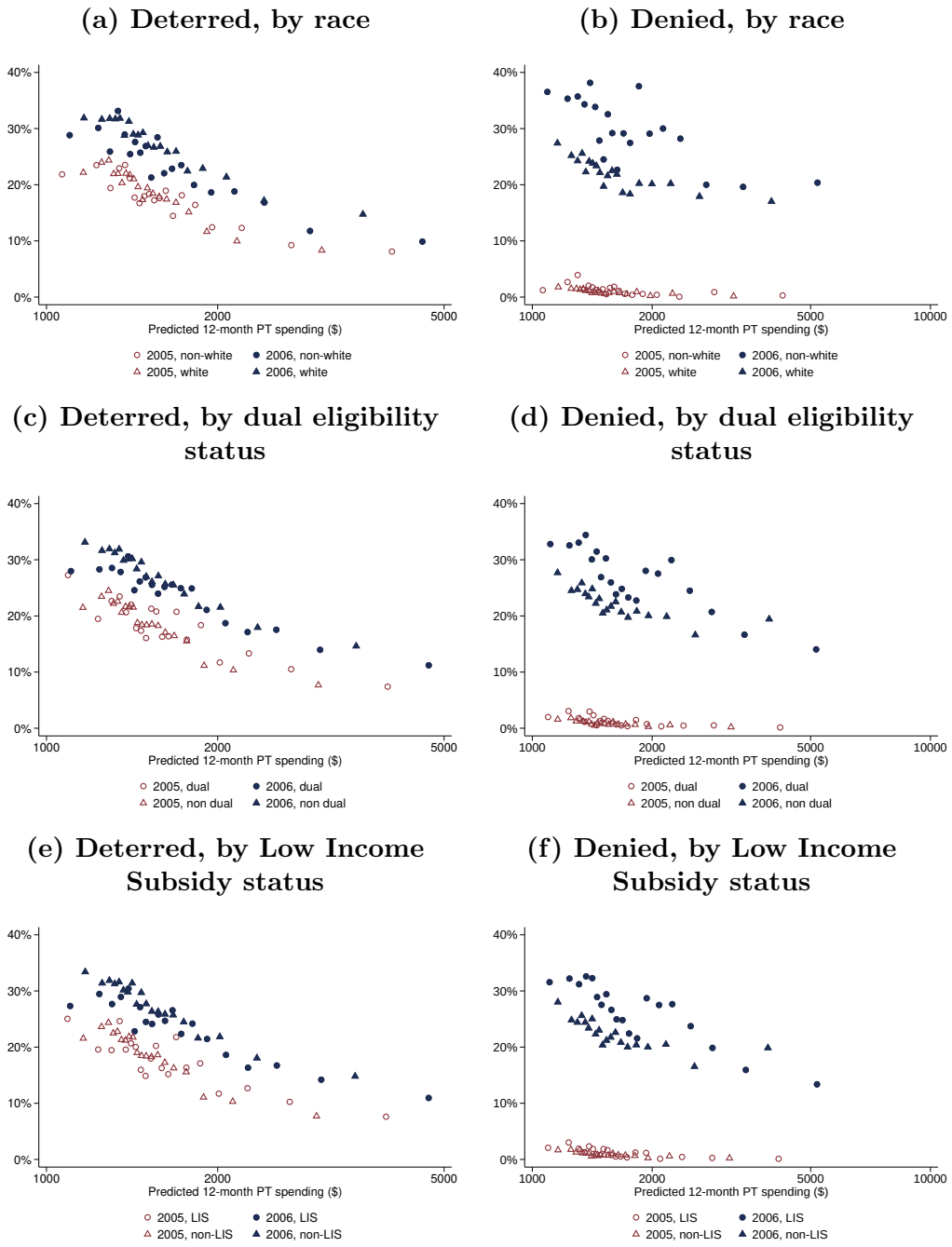
small providers in neither 2005 nor 2006. In contrast, panel (d) shows that this is only the case for denials prior to the introduction of the cap in 2005. Once the cap is introduced, patients who go to small providers are much less likely to be denied. The gap in denials between large and small providers exists even after controlling for patient predicted spending, indicating that it cannot be attributed to differences in patient composition across different providers. Instead, there are likely differences in behavior across providers that influence the denial rate, independent of patient need. At the same time, the results in panel (d) also show that the increased screening on the denial stage observed in panel (b) cannot be attributed to provider size – the slope of the correlation between patient need and Medicare denials still steepens, even after controlling for provider size. The estimates from Equation 6 in Table 2 columns (2), (4), and (6) confirm these patterns formally. Table A3 uses the estimates from Equation 6 to calculate the predicted denial rates among large and small providers for patients with the median levels of predicted need. Conditional on making an attempt, patients of smaller providers face an 83 percent (10.3 percentage point) higher denial rate.

By Patient Race and Income Any difference in denials between large and small providers introduced by the cap is itself concerning because it indicates that the cap is screening on something orthogonal to patient need. However, this would be even more concerning if different patient groups go to large and small providers, as the gap across providers would then translate into differences across patients. Figure B13 plots the correlation of the relationship between patient characteristics and an indicator variable for going to a small provider, among patients who approach the cap in 2006. Patients who go to small providers tend to be lower-income, as measured by average ZIP income and the share who are dually-eligible for Medicaid and who receive the Low Income Subsidy (LIS) for Part D. They are also more likely to be non-white.

These correlations between provider size and patient demographics motivate the analyses in Figure 8, which repeats the screening analysis in Figure 7, but splits by patient demographics. Table 3 shows the estimates of the coefficient on predicted spending, predicted spending interacted with a 2006 year dummy, and each respective demographic indicator as well as the indicator interacted with a 2006 year dummy. Columns (1)-(3) show that

among those who approach the cap, there is no difference in deterrence across patient race and income. Conversely, columns (4)-(6) show that among those who attempt, they are 27-34 percent more likely to be denied. On net, this translates into non-white and lower-income patients approaching the cap being 7-18 percent more likely to be stopped by the cap, holding fixed patient need (columns (7)-(9)). Table A3 shows that at the median level of predicted need among those who attempt, a non-white patient is 42 percent more likely to be denied than a white patient and dual-eligible/LIS patients are 33-35 percent more likely to be denied than non-dual and non-LIS patients.

Figure 8: Correlation between Patient Need and Deterrence or Denial by Patient Demographics, 2005-2006



This figure plots the relationship between attempt outcomes and patient need and provider size. "Deterred" is defined as the share of patients who approach the cap but do not attempt, and "denied" is defined as the share of patients who attempt but never make it past the cap. Panels (a) and (b) plot the relationship between log predicted 12-month PT spending and share deterred and denied in 2005 and 2006, split by patient race. Panels (c) and (d) plot the same relationship, split by patient dual eligibility status in the respective year. Panels (e) and (f) plot the same relationship, split by whether the patient receives the Part D Low Income Subsidy in 2006. Section 2.3 and Appendix Section E discuss the construction of the predicted PT spending measure and Section 2 describes the sample definition in further detail. Data: 20% Medicare Carrier claims and Master Beneficiary Summary File.

The emergence of the gap by provider size in Figure 7 and the strong correlation between provider size and patient demographics in Figure B13 are consistent with patient sorting across providers as a driver of the race and income gaps in Figure 8. However, an alternative explanation could be that the demographic gaps reflect across-group differences in Medicare decision-making about denials. This could be the case if, for example, Medicare's denial algorithm directly includes a race adjustment or indirectly inherits racial bias (Obermeyer et al., 2019). We evaluate this possible explanation by estimating Equations 5 and 6 with the inclusion of provider-year fixed effects, in Table 4. Once provider fixed effects are included, there are no within-provider disparities by race or income in neither deterrence nor denials. This indicates that the disparities see in Table 8 are indeed driven by patient sorting to different providers.[25] While this sorting did not lead to disparities in 2005, as shown by the null estimates on the coefficients for D_i themselves, once the cap was in place in 2006, which provider a patient sees becomes much more important in terms of their outcomes at the cap.

Taken together, the implication for welfare is mixed. While the cap may have improved the efficiency of spending by improving targeting, it also introduced horizontal inequity along a dimension unrelated to patient need. On the one hand, the cap improves targeting by medical necessity – patients with lower *ex ante* need see smaller increases in denials than patients with higher need from the cap. On the other hand, the cap introduces a substantial advantage for larger providers. Prior to the cap, two patients with similar medical need but who see differently-sized providers had similar chances of continuing care; after the cap, the patient who sees the smaller provider has a much higher chance of being denied. As a result, patients who tend to see smaller providers, like non-white and lower-income populations, are more likely to be stopped at the cap, even conditional on need.

[25] Furthermore, adding in 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 estimates for the "Predicted staffing × 2006" coefficients in Table 4, columns (5) and (9).

Table 3: Regression Results on Screening and Differences by Patient Demographics, 2005-2006

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Deterred			Denied					
	Approach			Attempt			Approach		
Predicted Spending	-17.3***	-17.5***	-17.4***	-1.98***	-2.03***	-2.04***	-19***	-19.2***	-19.2***
	(.713)	(.715)	(.729)	(.224)	(.232)	(.24)	(.741)	(.746)	(.763)
Predicted Spending × 2006	-3.31***	-3.17***	-3.13***	-10.3***	-10.6***	-10.6***	-7.56***	-7.71***	-7.67***
	(1.09)	(1.1)	(1.11)	(.935)	(.936)	(.941)	(1.25)	(1.25)	(1.26)
Non-white	.277			.287			.493		
	(.616)			(.231)			(.651)		
Non-white × 2006	-1.4			6.89***			3.62***		
	(1)			(1.15)			(1.22)		
Dual		.979*			.464*			1.34**	
		(.592)			(.238)			(.634)	
Dual × 2006		-1.54			5.8***			2.77**	
		(.999)			(1.07)			(1.21)	
Low-income subsidy			.276			.383*			.584
			(.579)			(.225)			(.613)
Low-income subsidy × 2006			-1.27			5.5***			2.88**
			(.977)			(1.05)			(1.18)
Outcome mean, 2005	18.7	18.7	18.7	1.4	1.4	1.4	19.9	19.9	19.8
Outcome mean, 2006	30.6	30.6	30.6	20.5	20.5	20.5	44.4	44.4	44.4
Week-year FE	X	X	X	X	X	X	X	X	X
Cluster	TIN	TIN	TIN	TIN	TIN	TIN	TIN	TIN	TIN
N. Clusters	11889	11889	11865	11057	11057	11032	11889	11889	11865
Observations	74024	74024	73298	56111	56111	55539	74024	74024	73298

This table presents the coefficients from estimating Equation 5 with log predicted PT spending as X_i (columns 1, 3, and 5) and Equation 6 with log predicted PT spending as X_i and an indicator for a patient's demographic characteristic D_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) . The coefficient β_1 is the estimate for "Predicted spending," the coefficient for β_2 is the estimate for "Predicted spending × 2006", the coefficient for β_3 is the estimate for a demographic indicator and β_4 is the estimate for the demographic indicator, interacted with an indicator for 2006. The regression is clustered at the provider (TIN-state) level. Columns 1-3 and 7-9 restrict to the sample of 2005-2006 patients who ever approach the cap and columns 4-6 restrict to the sample of patients who ever make an attempt to go past the cap, as defined in Section 2.3. Data: 20% Medicare Carrier claims.

Table 4: Regression Results on Screening and Differences by Patient Demographics with Provider FEs, 2005-2006

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Deterred				Denied							
	Approach				Attempt				Approach			
Predicted Spending	-13.5***	-13.5***	-13.5***	-13.5***	-.647***	-.646***	-.649***	-.651***	-14.1***	-14.1***	-14.2***	-14.1***
	(.846)	(.846)	(.845)	(.86)	(.205)	(.205)	(.206)	(.209)	(.862)	(.861)	(.86)	(.88)
Predicted Spending × 2006	-1.08	-1.09	-1.06	-1.09	-5.9***	-5.9***	-5.92***	-5.93***	-3.72***	-3.72***	-3.71***	-3.75***
	(1.27)	(1.27)	(1.26)	(1.27)	(.81)	(.81)	(.81)	(.811)	(1.32)	(1.32)	(1.32)	(1.33)
Non-white		.889				-.36				.708		
		(.793)				(.26)				(.814)		
Non-white × 2006		-1.93				1.13				-1.31		
		(1.2)				(.797)				(1.21)		
Dual			.823				.0613				.905	
			(.768)				(.303)				(.805)	
Dual × 2006			-.423				.93				.315	
			(1.2)				(.818)				(1.25)	
Low-income subsidy				.134				-.0145				.177
				(.749)				(.298)				(.781)
Low-income subsidy × 2006				-.408				1.26				.648
				(1.17)				(.821)				(1.22)
Outcome mean, 2005	18.4	18.4	18.4	18.3	1.3	1.3	1.3	1.2	19.4	19.4	19.4	19.4
Outcome mean, 2006	29.8	29.8	29.8	29.8	17.7	17.7	17.7	17.7	42.7	42.7	42.7	42.7
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	TIN	TIN	TIN	TIN	TIN	TIN	TIN	TIN	TIN	TIN	TIN	TIN
N. Clusters	9244	9244	9244	9219	7987	7987	7987	7960	9244	9244	9244	9219
Observations	68254	68254	68254	67539	50377	50377	50377	49814	68254	68254	68254	67539

This table presents the coefficients from estimating Equation 5 with log predicted PT spending as X_i (columns 1, 5, and 9) with provider fixed effects, and Equation 6 with log predicted PT spending as X_i and an indicator for a patient's demographic characteristic D_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) with provider fixed effects . The coefficient β_1 is the estimate for "Predicted spending," the coefficient for β_2 is the estimate for "Predicted spending × 2006", the coefficient for β_3 is the estimate for a demographic indicator and β_4 is the estimate for the demographic indicator, interacted with an indicator for 2006. The regression is clustered at the provider (TIN-state) level. Columns 1-4 and 9-12 restrict to the sample of 2005-2006 patients who ever approach the cap and columns 5-8 restrict to the sample of patients who ever make an attempt to go past the cap, as defined in Section 2.3. Data: 20% Medicare Carrier claims.

5 Mechanism for Differences By Provider Size

5.1 The Role of Documentation

The results in Tables 2, 3, and 4 show that there is substantial variation in denials across providers, which in turn translates into variation across patient race and income groups, that could not be explained by patients' clinical need. One possible source of this variation is in providers' documentation practices. While documentation was not strictly required for an exception to be approved, CMS placed significant emphasis on documentation in its communications about getting approvals for exceptions to the therapy cap (Centers for Medicare and Medicaid, 2006c). In this section, we argue that the main driver of these across-provider differences in denials is variation in compliance with *documentation* requirements across large and small providers.

We first show evidence that documentation is highly related to denials both in the cross-section and in over time. Table 5 shows the estimates of a regression between an indicator for the use of documentation on an attempt in 2006, as measured by the use of the documentation modifier code on any claim in the week of an attempt, and the likelihood that the attempt is approved. There is a strong positive correlation between the two – 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 (column (2)), patient predicted health (column (3)), and provider size (column (4)), consistent with the effect being due to the documentation itself as opposed to patient or provider characteristics correlated with documentation use.

Table 5: Regression of Documentation on Approvals on Cap Attempts, 2006

	(1)	(2)	(3)	(4)
	Share approved			
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 KX	.018	.019	.024	.032
R^2, with KX	.068	.067	.074	.075
Outcome mean	46.878	47.022	46.247	47.022
Patient demographics	X	X		X
Patient health		X		X
Provider size			X	X

This table reports the results of regressions on approvals in 2006, conditional on their provider attempting to provide care past the cap. The outcome variable is the share of attempts where the patient faces no denials in that week, where an attempt is defined as in Section 2.3: a week in which a patient's cumulative paid amount is below the cap at the beginning of the week, and their cumulative paid or billed amount is above the cap at the end of the 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 "Use 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. Standard errors (in parentheses) are clustered at the provider level. Data: 20% Medicare Carrier claims and Master Beneficiary Summary File.

Documentation and denials are also closely related to each other over time. Figure 9 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 at zero because the modifier code was not introduced yet. Only about a quarter of attempts included documentation at the beginning of 2006, and this rises to about three-quarters by 2007. Strikingly, the rate of claim approvals moves in lock-step with the increasing documentation of medical need. This co-movement strongly suggests that the

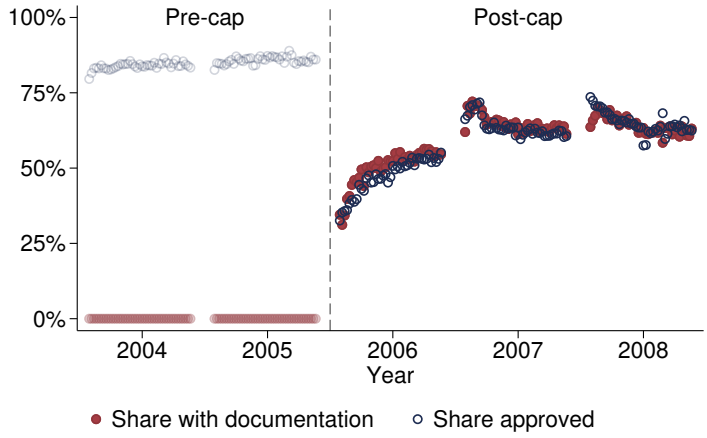
increase in documentation underlies the increase in approvals.

In Figure B15, we investigate whether the change in approval rates over time could reflect changes in claim-level strategic *billing* behavior. To measure this, we train a machine-learning model to predict approval rates on claims associated with attempts that includes rich line-level diagnosis, procedure, and modifier codes, as well as patient characteristics and prior utilization. This model can capture high-dimensional interactions between claim-level characteristics that could affect the approval rate and would not be captured in our patient-level measure of predicted need. In order to capture strategic billing behavior that is *independent* of documentation usage, we use our model to create a predicted approval rate under the assuming that every attempt uses documentation. The time series of this predicted approval rate 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.

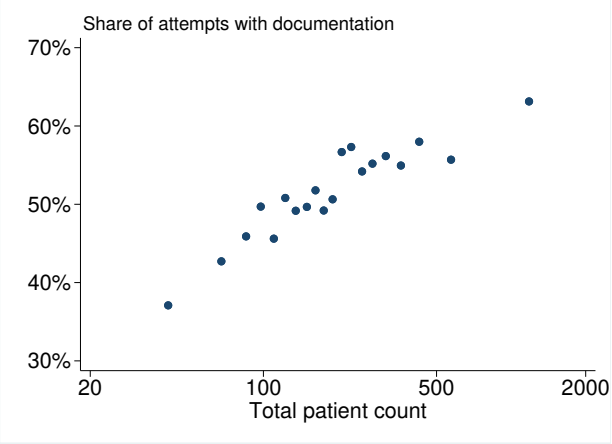
Finally, we explore the relationship between provider size and documentation use. The binscatters in Figure 9 panels (b) and (c) show that provider size is highly correlated with both documentation use and approvals. Providers in the top quartile of size, who see about 1500 Medicare patients from 2006-2008, are 22 percent (11.8 percentage points) more likely to use documentation on attempts in 2006 than those in the bottom quartile, who see 575 patients (panel a). This translates into a sizeable differences in approvals by size – attempts by providers in the top quartile of size are 48 percent (16.1 percentage points) more likely to be approved than those in the bottom quartile (panel b).

Figure 9: Documentation and Approval on Cap Attempts

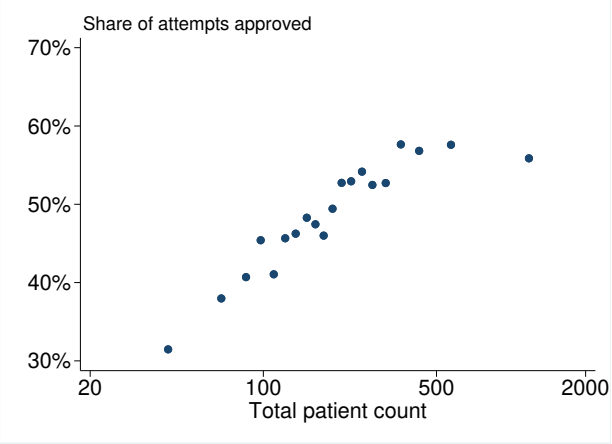
(a) Time Series, 2006-2008



(b) Documentation Use vs. Provider Size, 2006



(c) Approvals vs. Provider Size, 2006



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. "Share with documentation" is the share of attempts in a given calendar week 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. Attempts are defined as in Section 2.3: weeks in which a patient's cumulative paid amount is below the cap at the beginning of the week, and their cumulative paid or billed amount is above the cap at the end of the week. Provider size is measured as a TIN-state's 2006-2008 Medicare patient count. Data: 20% Medicare Carrier claims.

5.2 Link between Provider Size and Documentation: Learning

Decomposition of Provider Size Advantage Figure 9 panels (b) and (c) show large providers are more likely to use documentation and face fewer denials at the cap. Furthermore, our results in Figure 7 show that this pattern is independent of a patient's underlying need. We next explore *why* larger providers have better compliance with documentation requirements and thus fare better in the face of the cap.

Documentation practices can vary across providers of different sizes for a variety of reasons. On the one hand, large and small providers could differ at baseline: they could have differing investments in technology to document medical need, like EHR and billing software. Likewise, larger providers may have invested more in human capital by hiring more specialized staff to handle billing. Furthermore, the correlation between size and documentation use could also reflect an omitted factor that is correlated with both size and documentation use, like ownership status or geographic location.

On the other hand, Figure 9 suggests an alternative mechanism for large-provider advantage: economies of learning. In particular, there is a rapid increase in documentation and approvals throughout the first year of the policy, suggesting that providers learned a significant amount about avoiding denials during this period. Insofar as providers learn through their experience what documentation and billing practices are likely to be approved—i.e., "learning-by-doing"—then large providers derive an advantage from moving up the learning curve more quickly. This is demonstrated in panel (a), which depicts the amount of experience that small and large providers accumulate over time. As expected, small providers—those below the median—accumulate experience with the cap much more slowly than large providers. By the end of the first year, approximately 60% 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. By the end of 2008, essentially all large providers have over 50 patients worth of experience, while only about 40% of small providers have reached this mark.

However, learning need not imply a large-provider advantage. If information about successful documentation and billing practices naturally permeates the industry over time, then

this "industry-wide learning" would not generate advantage through economies of learning. For example, Medicare and industry groups could have been disseminating information throughout 2006 to large and small providers alike on how effective documentation dramatically reduces the likelihood of denial.

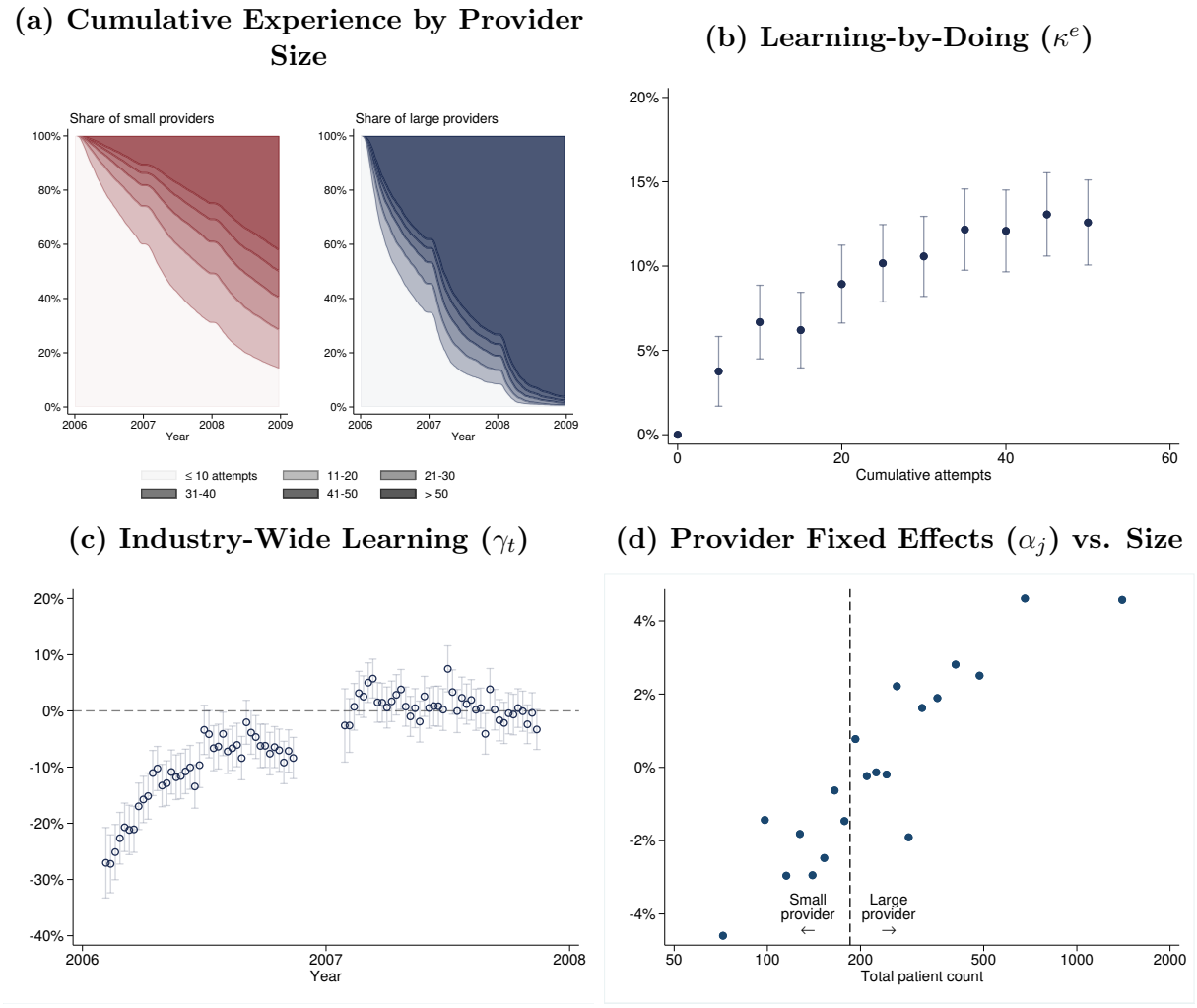
We next decompose the variation in denials into learning-by-doing—which give large providers a short-term advantage through faster accumulation of experience—and persistent size-based advantages that do not attenuate with time. We estimate a linear model of approval or documentation use Y_i for patient i' receiving care from provider j that is attempting to bill above the cap in week t:

$$Y_i = \underbrace{\sum_{e(j(i),t(i))=1}^{50} \kappa^e}_{\text{Learning-By-Doing}} + \underbrace{Week_{t(i)}}_{\text{Industry-Wide Learning}} + \underbrace{\alpha_{j(i)}}_{\text{Provider FE}} + \underbrace{\beta X_i}_{\text{Patient Controls}} + \varepsilon_i \tag{7}$$

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 approach the cap.[26] The second is, $\gamma_{t(i)}$, which captures the accumulated effects of industry-wide learning for providers by period $t(i)$, i.e., the amount of learning over time that cannot be explained by providers' individual experiences. The last is α_j, a fixed-effect for the persistent differences in provider denial rates. This captures variation in providers aptitude for billing that may result from investments or other provider characteristics.

[26]Formally, $e(j(i), t(i))L = |\{\iota : j(\iota) = j(i), t(\iota) < t(i)\}|$ is the count of patients who approached the cap with provider j before i approached the cap with j.

Figure 10: Decomposition of Size Advantage on Approvals of Attempts



This figure characterizes differences in cap attempt outcomes by provider size and experience in 2006-2008 using data and regression estimates from Equation (7). Panel (a) plots the share of big and small providers and their cumulative attempts over time. "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. Panel (c) provides estimates of industry-wide learning for 2006 and 2007, normalized by the equivalent-week coefficients for 2008. Panel (d) plots the provider fixed effects against log provider patient count. Section 2.3 describes the definition of cap attempts in further detail. Data: 20% Medicare Carrier claims.

Figure 10 panels (b)-(d) present the estimates from Equation 7 to decompose the size advantage in approvals. Panel (b) shows the estimated economies-of-learning curve and plots κ^e for $e \in [1, 50]$. The estimates suggest a significant amount of learning-by-doing, even when controlling for industry-wide learning and persistent variation across providers.

The implied advantage from being farther down the economies of learning curve can be quite large, with a veteran provider with more than 50 patients worth of experience having a 14 percentage point advantage over a complete novice. In the context of panel (a), this implies that economies of learning likely contributed substantially to large providers' advantage, especially in the first few years before the experience of smaller providers started to reach significantly diminishing marginal returns.

Panel (c) shows our estimates of industry-wide learning for the first two years after the introduction of the soft cap. These suggest significant industry-wide learning that occurred throughout the first year or so with no new industry-wide learning occurring throughout the second year. This industry-wide learning explains an approximately 10 percentage point increase in industry-wide approval rates throughout the first year or so.

Note that we do not present estimates of $\gamma_{t(i)}$ directly, since these are biased by selection in patients that reach the cap over time. In particular, we may be concerned that patients who reach the cap earlier in the calendar year have greater medical need. The estimates of $\gamma_{t(i)}$ are consistent with such within-calendar-year changes in patient composition (Appendix Figure B16). To address these within-year composition changes, Figure 10 panel (c) presents estimates of the $\gamma_{t(i)}$ from 2006 and 2007 that difference out the coefficients from the analogous weeks in 2008. Insofar as new industry-wide learning is not accumulating in 2008, then these these 2008 coefficients purely represent within-year selection, and differencing them out helps isolate the effect of industry-wide learning. Insofar as industry-wide learning is continuing in 2008, then the differenced estimates represent the additional industry-wide learning occurring in 2006 and 2007 relative to 2008. That industry-wide learning would have concluded more than two years into the policy seems plausible, especially given that our estimates suggest little to no learning even by 2007 relative to 2008.

Finally, panel (d) plots a binscatter of provider fixed effects against provider size. These suggest a significant amount of variation in large-provider advantage that persists even when accounting for both learning-by-doing and industry-wide learning. Specifically, the largest providers appear to have an approximately 6 percentage point persistent advantage relative to the smallest providers. This advantage may represent physical or human capital investments from large providers that are difficult for small providers to replicate.

Taken together, Figure 10 shows that large providers face fewer denials both due to economies of learning as well as a persistent advantage that is likely due to investments in billing. There is also a significant amount of industry-wide learning that occurred after the policy. Insofar as regulators want to avoid spending limits conferring advantages to large firms, they might consider simplifying or clarifying exceptions processes such that learning can occur industry-wide rather than through individual provider experience or investments too costly for smaller firms.

Sharp Evidence of Learning-By-Doing As a final piece of evidence on the economies of learning, we look at how provider behavior evolves around events which should be associated with learning. We focus on the first time that the provider successfully reverses a denial by adding documentation. In particular, these are defined as weeks 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.

Following Cengiz et al. (2019), we use a stacked event study method. Each stack is centered around a given provider's "learning event," and we look at that provider's 5 attempt-weeks before and after their event with *other* patients. 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 in the 5 attempt-weeks before or after the focal week.[27]. A given attempt can appear in multiple stacks as part of the control group, but can only appear in one stack.

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

$$Y_{igt} = \sum_{a(i,g)=-5}^{5} 1(RelWeek_{a(i,g)} = a) \times Treat_{j(i),g} + \underbrace{\alpha_{j(i)}}_{\text{Provider FE}} + \underbrace{\gamma_g}_{\text{Stacked Group FE}} + \underbrace{\beta X_i}_{\text{Patient Controls}} + \underbrace{RelWeek_{a(i,g)}}_{\text{Week Relative to Attempt FE}} + \underbrace{Week_t}_{\text{Calendar Week FE}} + \varepsilon_{igt}. \quad (8)$$

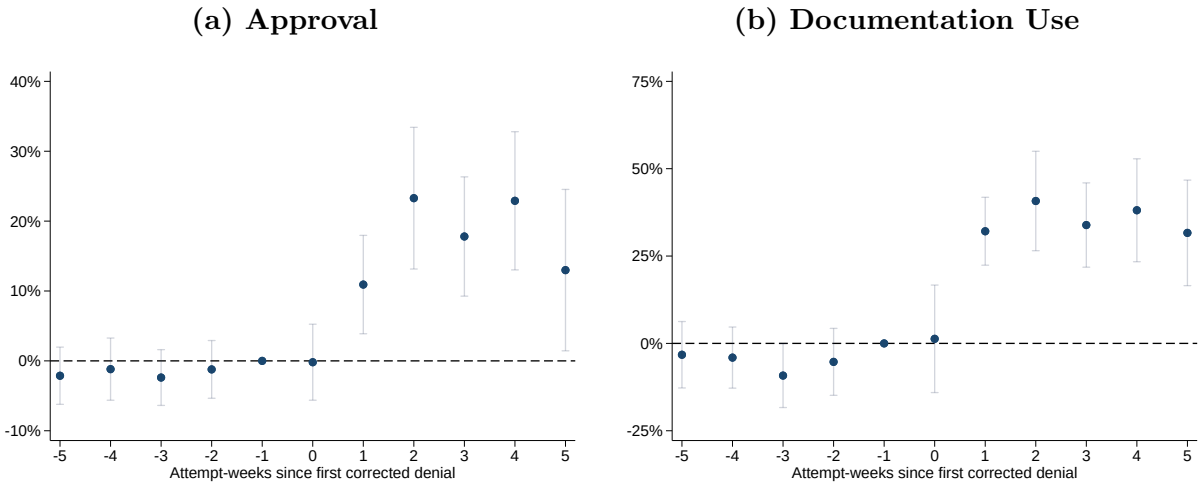
[27]In other words, they are either not treated in the entire period (because they are not-yet-treated or they are never-treated) or treated in the entire period.

$Treat_{j,g}$ is an indicator variable for whether the attempt with patient i is associated with provider j's "learning event." λ_j is a provider fixed effect and γ_g is a stacked group fixed effect. We also control for X_i, the predicted need associated with the patient. Finally, $Week_t$ is the calendar week associated with patient i's attempt and $RelWeek_a$ is the attempt-week (between -5 and 5) *relative* to the learning event that defines group g; these can be separately identified because not all providers attempt in every week. Results are clustered at the provider-level to account for repeated observations across stacks.

Figure 11 plots the coefficients from Equation 8 where in panel (a) Y_{igt} is an indicator whether the attempt for patient i in week t was approved, and panel (b) is an indicator for whether the provider used documentation in their attempt. Prior to the learning event, providers were not on a differential trend relative to control providers who also made an attempt in the same week as the learning event. However, after the provider experiences the reversed denial, they are consistently more likely to include documentation on future attempts and more likely to be approved.

These results provide sharp evidence of within-provider behavior change. These event study results alone could be interpreted as being driven by either learning or investment – in the latter case, the "learning event" could just be the week that the provider installs software to collect documentation. However, if the event study only reflected investment, we would not also see the learning-by-doing curve depicted in Figure 10 panel (b). Together, these pieces of evidence point to within-provider learning as explaining some part of the decrease in denial rates over time.

Figure 11: Outcomes on Attempts Around Provider Learning Event, 2006



This figure plots the coefficients from a stacked event study specification around a provider's "learning event" in Equation 8. The outcome variable is (a) the share of attempts approved and (b) the share of attempts with documentation in a given attempt-week on beneficiaries other than the one associated with the learning event. The learning event is defined as a patient for whom a provider originally attempted without documentation and was denied at the cap, and then attempted again with documentation and got past the cap. An attempt-week is a week in which the provider made at least one attempt. 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 and Master Beneficiary Summary Files.

6 Conclusion

This paper studies how soft spending limits work by examining a spending limit imposed on healthcare providers by Medicare for physical therapy. We using bunching methods to find that the soft "therapy cap" reduced overall spending, though much less than when compared to a case study of a hard cap that limits all spending above the cap. We look for direct effects of these savings on patient health use an identification strategy that leverages the annual nature of the cap, which resets every January. We show that the cap is more binding for patients who start PT earlier in the year than those who start later, and that this pattern only emerges in 2006. Looking to health effects, we find precise null effects on PT substitutes like opioid use, pain procedures, and orthopedic surgeries, as well as on ED

visits and inpatient or skilled nursing stays.

We then explore the effect of the cap on the targeting of spending by characterizing which patients are stopped at the cap, and why. Comparing how the correlation between receiving care past the cap and patient need changes before and after the cap is instituted, we find that while spending decreases for patients across the board, the reductions in spending are larger for lower-value spending. Thus, the cap improves the targeting of spending. Using a novel feature of healthcare claims data which allows us to observe both successful and unsuccessful attempts by providers to bypass the cap, we separate the bunching response into a deterrence channel – providers who respond to the cap by stopping just short of it without making an attempt – and the denial channel – providers who make an attempt but are forced to stop by Medicare. The majority of marginal patients are stopped by denial, rather than deterrence.

We also find that the improvements in targeting stem entirely from Medicare being more discerning in issuing denials, rather than providers being more selective in who they treat. We also find evidence that the cap introduces inequities by provider size which did not exist before: patients who go to small providers were much more likely to be stopped by the cap, as their providers faced more Medicare denials. This disparity holds even after accounting for patient need. As lower-income and minority patients tend to see smaller providers, the disparities across providers translate into large gaps across patient race and income.

We then dig in to the mechanisms underlying these differences by provider size. All else equal, claims are much less likely to be denied if providers submit documentation with their attempt to go past the cap, and larger providers are much more likely to submit documentation with their attempts. We decompose the large provider advantage into the component driven by experience ("learning by doing"), secular changes over time, and fixed provider-specific effects. We find that providers become more successful at navigating the cap with experience, so some of the disparities between large and small providers are explained simply by smaller providers having less experience. However, large providers still enjoy an advantage up to two years later. We also provide sharp evidence from an event study design showing that providers learn from billing mistakes – after correcting their first cap denial, providers are much more likely to use documentation and receive an approval on subsequent

attempts.

Taken together, our findings have mixed implications on the efficacy of soft spending limits in resolving the principal-agent issues that arise when expenditure decisions are delegated. On the one hand, we show that imposing such a limit does indeed improve the efficiency of spending – spending limits allow principals to strike a balance between reducing overall spending while still allowing high-value spending to occur. On the other hand, we also provide evidence that these limits introduce perhaps undesirable horizontal inequity – say, by differences in how well agents can handle the administrative ordeal required to bypass the limit.

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.

American Physical Therapy Association, "MedPAC Releases Revised Medicare 'Payment Basics'," Technical Report October 2020.

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.

Brot-Goldberg, Zarek C., Amitabh Chandra, Benjamin R. Handel, and Jonathan T. Kolstad, "What does a Deductible Do? The Impact of Cost-Sharing on Health Care Prices, Quantities, and Spending Dynamics*," *The Quarterly Journal of Economics*, August 2017, *132* (3), 1261–1318.

_ , Samantha Burn, Timothy Layton, and Boris Vabson, "Rationing Medicine Through Bureaucracy: Authorization Restrictions in Medicare," January 2023.

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.

Centers for Medicare and Medicaid, "Outpatient Therapy Caps: Exceptions Process Required by the DRA," February 2006.

_ , "Therapy Caps Exception Process," Technical Report Transmittal 855 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 June 2006.

_ , "Therapy Caps and Advance Beneficiary Notice of Noncoverage (ABN), Form CMS-R-131, FAQs April 2013," April 2013.

_ , "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)," 2020.

_ , "Medicare Benefit Policy Manual. Chapter 15 – Covered Medical and Other Health Services," 2024.

Chandra, Amitabh, Jonathan Gruber, and Robin McKnight, "Patient Cost-Sharing and Hospitalization Offsets in the Elderly," *American Economic Review*, March 2010, *100* (1), 193–213.

Code of Federal Regulations, "Approval of the justification," 2020.

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.

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 J., Riley J. League, Jetson Leder-Luis, Ryan C. McDevitt, and James W. Roberts, "Ambulance Taxis: The Impact of Regulation and Litigation on Health Care Fraud," November 2021.

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.

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.

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," 2022.

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., "Does It Pay to Know Prices in Health Care?," *American Economic Journal: Economic Policy*, February 2017, *9* (1), 154–179.

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

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

Manning, Willard G., Joseph P. Newhouse, Naihua Duan, Emmett B. Keeler, and Arleen Leibowitz, "Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment," *The American Economic Review*, 1987, *77* (3), 251–277. Publisher: American Economic Association.

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.

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

O'Malley, A. James, Thomas A. Bubolz, and Jonathan S. Skinner, "The Diffusion of Health Care Fraud: A Network Analysis," Working Paper w28560, National Bureau of Economic Research March 2021.

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

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

US General Accounting Office, "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.

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 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)
	Deterred		Denied			
	Approach		Attempt		Approach	
Log age	2.42*	2.46*	-.467	-.481	1.96	1.99
	(1.28)	(1.28)	(.522)	(.523)	(1.33)	(1.33)
Log age × 2006	-.53	-.61	-11.6***	-10.6***	-7.96***	-7.43***
	(2.21)	(2.21)	(2.05)	(2.01)	(2.59)	(2.57)
Small provider		-1.19**		.294*		-.956*
		(.536)		(.176)		(.563)
Small provider × 2006		.707		10***		7.47***
		(.909)		(.849)		(1.11)
Outcome mean, 2005	18.7	18.7	1.4	1.4	19.9	19.9
Outcome mean, 2006	30.6	30.6	20.5	20.5	44.4	44.4
Week-year FE	X	X	X	X	X	X
Cluster	TIN	TIN	TIN	TIN	TIN	TIN
N. Clusters	11889	11889	11057	11057	11889	11889
Observations	74024	74024	56111	56111	74024	74024

This table presents the coefficients from estimating Equation 5 with log patient age as X_i (columns 1, 3, and 5) and Equation 6 with log patient age as X_i and an indicator for whether a patient goes to an above-median-size provider as D_i (columns 2, 4, and 6). The coefficient β_1 is the estimate for "Predicted spending," the coefficient for β_2 is the estimate for "Predicted spending × 2006", the coefficient for β_3 is the estimate for "Small provider" and β_4 is the estimate for "Small provider × 2006." The regression is clustered at the provider (TIN-state) level. Provider size is measured as a TIN-state's 2006-2008 Medicare patient count and "small provider" denotes a provider with below-median patient count. Columns 1-2 and 5-6 restrict to the sample of 2005-2006 patients who ever approach the cap and columns 3-4 restrict to the sample of patients who ever make an attempt to go past the cap, as defined in Section 2.3. Data: 20% Medicare Carrier claims.

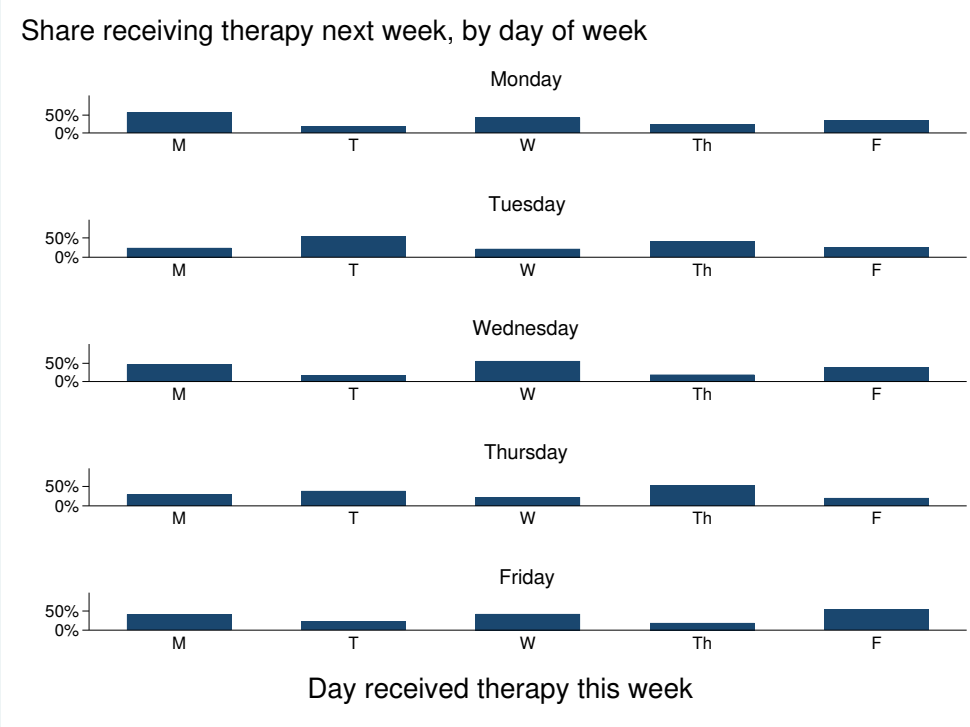
Table A3: Differences by Race and Income on Denials for Median Patient

	(1)	(2)
	Denied	
	Attempt	Approach
Median predicted spending in 2006	1509	1483
Median week of year	24	25
Predicted rates		
Small provider	12.4	38.2
Small - large provider difference	10.3 (83%)	7.8 (20%)
White	16.5	40.9
Non-white - white difference	6.9 (42%)	3.6 (9%)
Non-dual eligible	16.5	40.8
Dual - non-dual difference	5.8 (35%)	2.8 (7%)
Non-LIS	16.5	40.9
LIS - non-LIS difference	5.5 (33%)	2.9 (7%)

This table presents the implied denial rate in 2006, holding fixed patient predicted spending and the week of approach and attempt, and comparing across race and income groups. The predicted denial rate is calculated using the coefficients from Table 3. Column 1 conditions on patients who make an attempt and column 2 conditions on patients who make an approach, as defined in Section 2.3. Data: 20% Medicare Carrier claims.

B Appendix Figures

Figure B1: Share Receiving Therapy on Each Day of Week

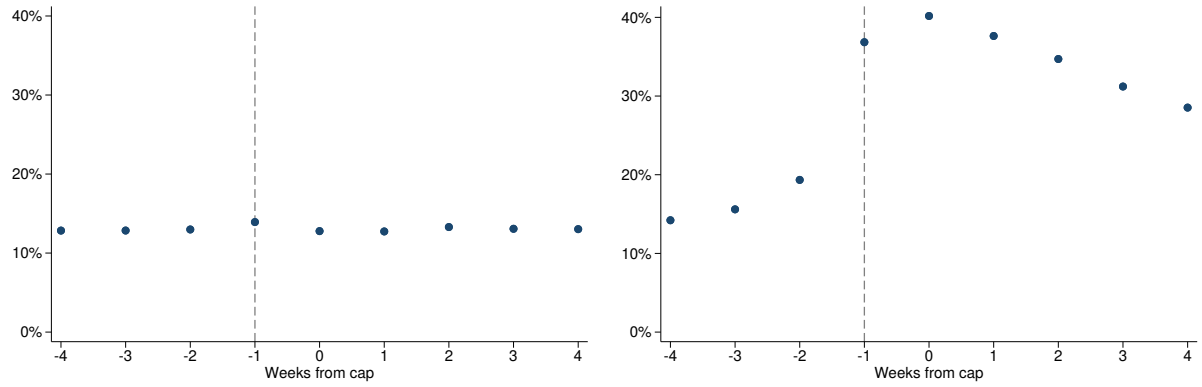


This plot graphs the share of PT visits on each day of the week, given the day of therapy in the previous week. Data: 20% Medicare Carrier claims.

Figure B2: Denial Rates Pre- and Post-Reform

(a) Pre-Reform, 2005

(b) Post-Reform, 2006

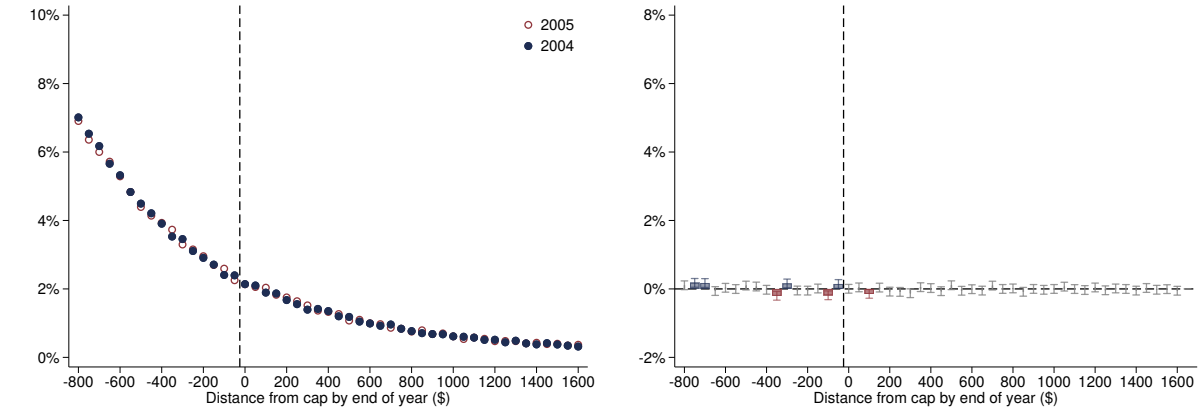


This figure plots the average denial rate for PT claims depending on how far the patient is from the therapy cap, where the denial rate is the share of claims with at least one denied line. Panel (a) graphs the denial rates relative to a (placebo) 2006 cap in 2005, and panel (b) graphs the denial rates 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. Denial rate is defined as the share of lines in an attempt-week that are denied. Data: 20% Medicare Carrier claims.

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

(a) Distribution of Spending

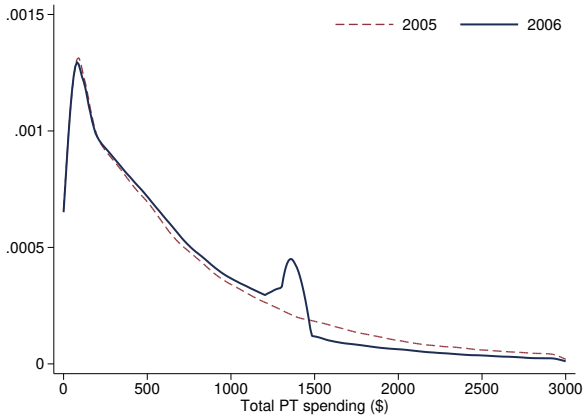
(b) Difference between 2005 and 2004



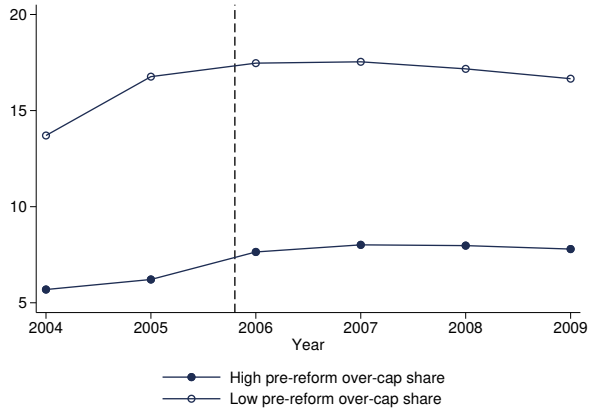
This figure plots the (a) distributions of end-of-year physical therapy spending around the cap in 2004 and 2005 and (b) the difference in the distributions between 2004 to 2005. Distance from cap is calculated in bins of $50 relative to the 2006 cap and shares are calculated as the share of patients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

Figure B4: Extensive Margin Responses

(a) Distribution of PT Spending, 2005-2006



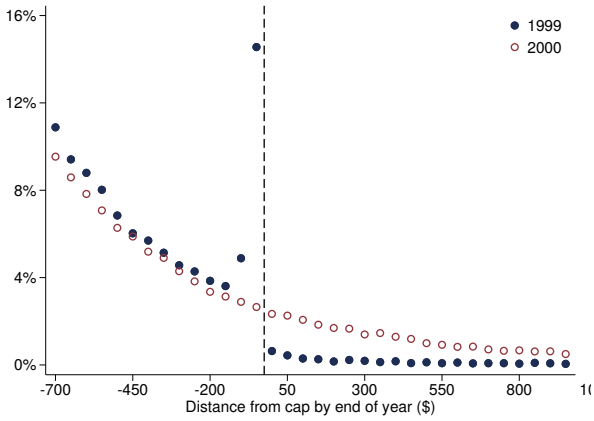
(b) Mean beneficiaries per provider, by 2005 over cap share



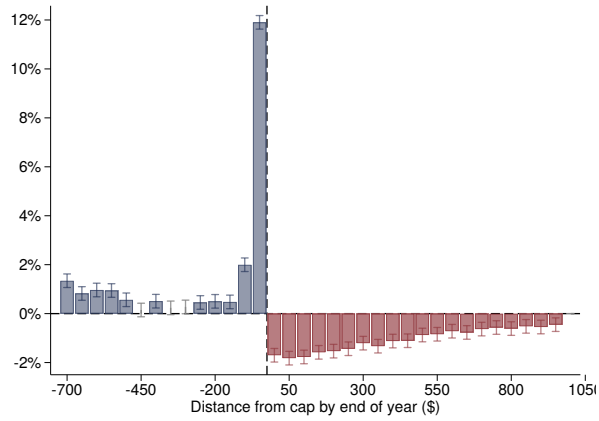
This figure characterizes potential extensive margin responses to the 2006 therapy cap. Panel (a) plots a kernel density of Medicare physical therapy spending in 2005 and 2006 up to $3000. Since Medicare pays for 80 percent of allowed charges and patients are responsible for the remaining 20 percent, 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

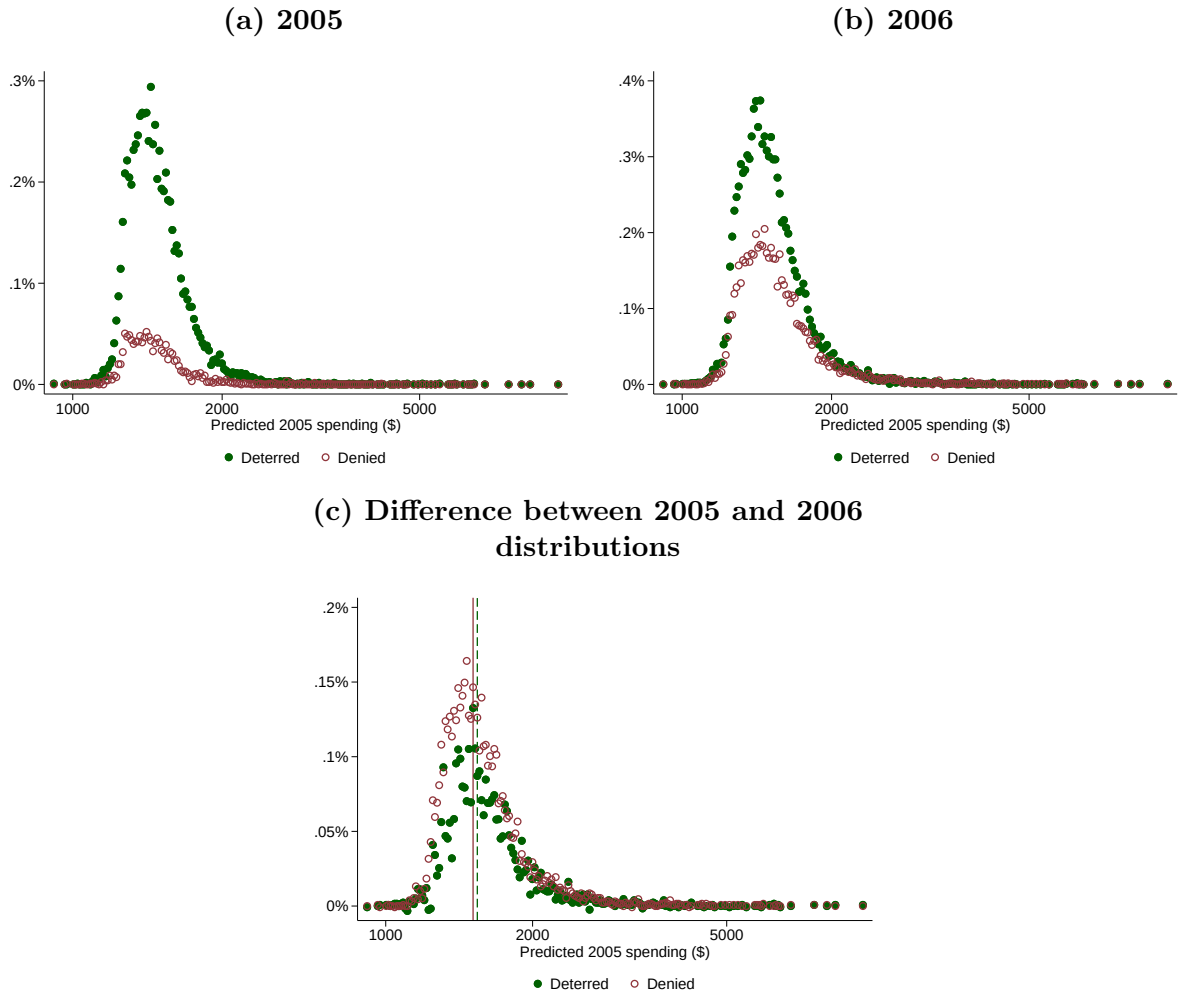


(b) Difference between 1999 and 2000



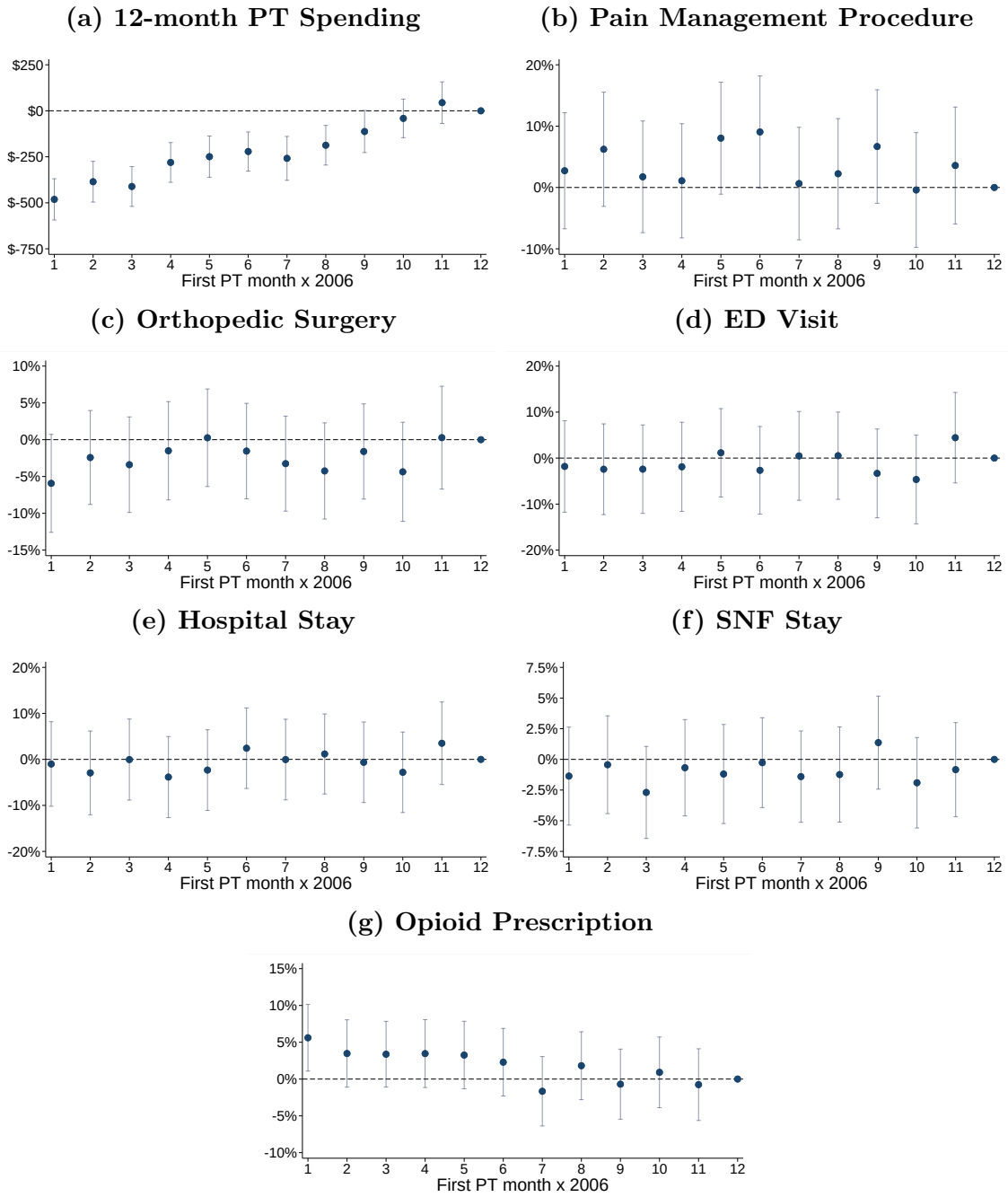
This figure plots the (a) distributions of end-of-year physical therapy spending around the cap in 1999 and 2000 and (b) the difference in the distributions between 1999 and 2000. Distance from cap is calculated in bins of $50 relative to the 1999 cap (inflation-adjusted to 2005 dollars) and shares are 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



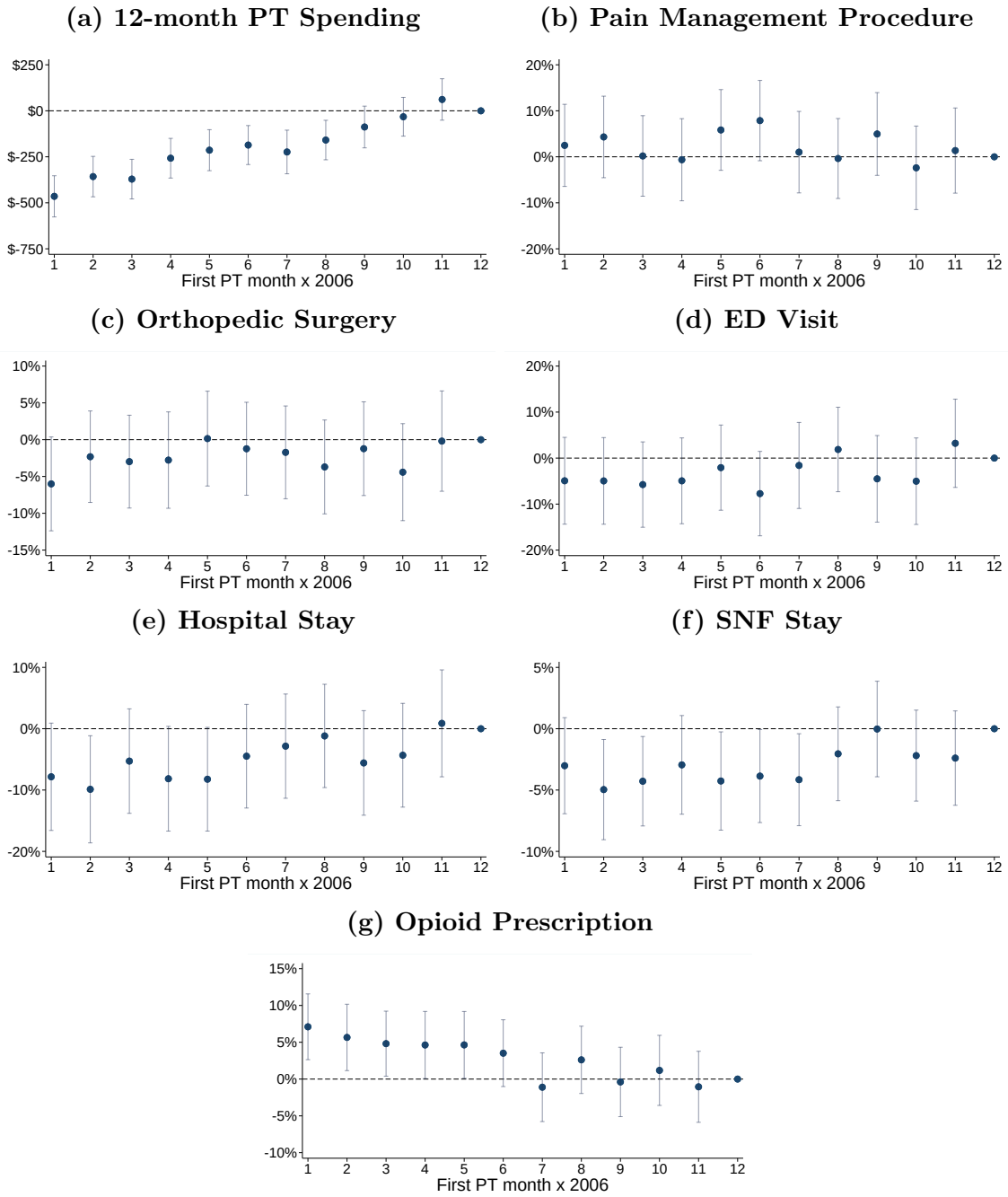
This figure plots the distribution predicted 2005 PT spending for "deterred" patients, defined as those who stop care one week before the therapy cap without making an attempt, and "denied" patients, defined as those who stop care before the therapy cap after a denied attempt. Panel (a) plots the distribution in 2005 at the placebo 2006 cap, panel (b) plots the distribution 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 E describe the predicted PT spending measure in further detail. Data: 20% Medicare Carrier claims.

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



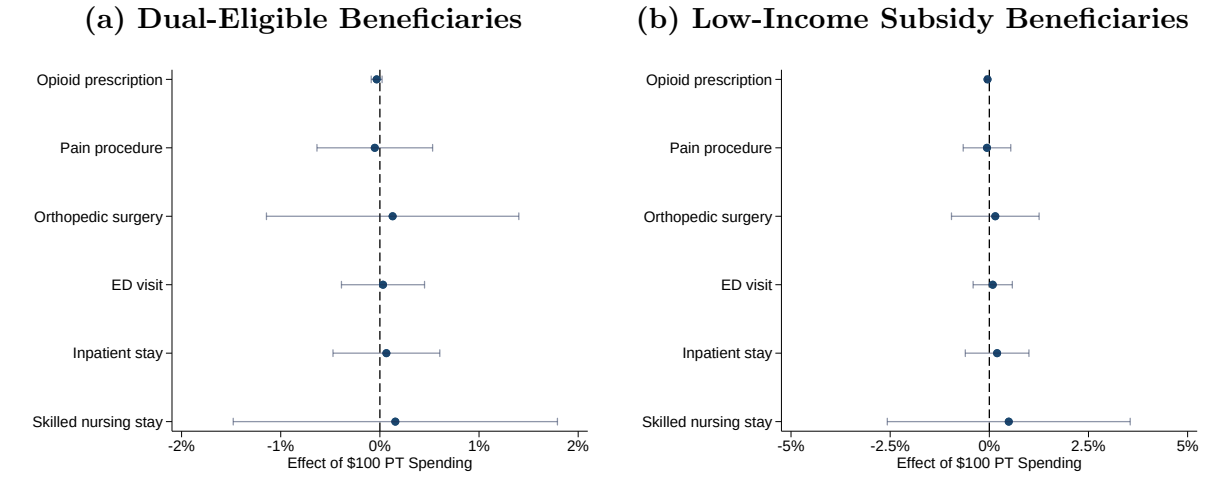
This figure plots the coefficient θ_m, which denotes the interaction between an indicator for 2006 and an indicator for month of first PT, from Equation 2. Panel (a) plots the coefficients on 12-month PT spending ($). Panel (b) plots the coefficients on an indicator for pain management procedures, panel (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 SNF 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 F describes how the health outcome measures are defined. Sample is restricted to beneficiaries with less than $200 in 12-month PT spending who are dual-eligible for Medicare and Medicaid. Data: 20% Medicare Carrier, Outpatient, MEDPAR, and Part D claims.

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



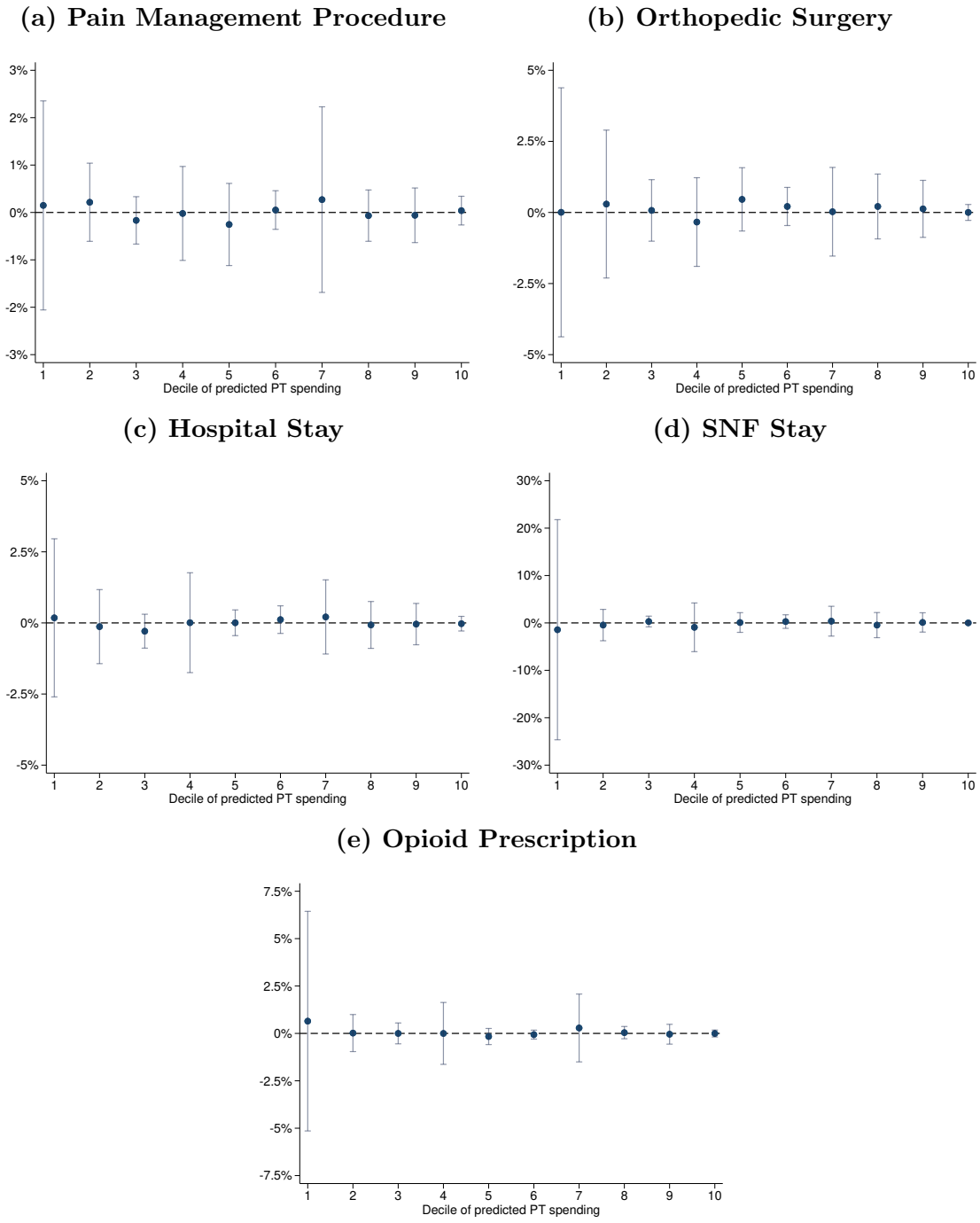
This figure plots the coefficient θ_m, which denotes the interaction between an indicator for 2006 and an indicator for month of first PT, from Equation 2. Panel (a) plots the coefficients on 12-month PT spending ($). Panel (b) plots the coefficients on an indicator for pain management procedures, panel (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 SNF 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 F describes how the health outcome measures are defined. Sample is restricted to beneficiaries with less than $200 in 12-month PT spending who receive the Low Income Subsidy in 2006. 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, from Equation 4. Panel (a) subsets to dual-eligible beneficiaries and panel (b) subsets to beneficiaries receiving LIS. Since LIS is not available until 2006, eligibility for the 2005 sample is based on their 2006 eligibility. 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 and are scaled in percent terms relative to the sample average. Section F describes how the health outcome measures are defined. Sample is restricted to beneficiaries with less 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.

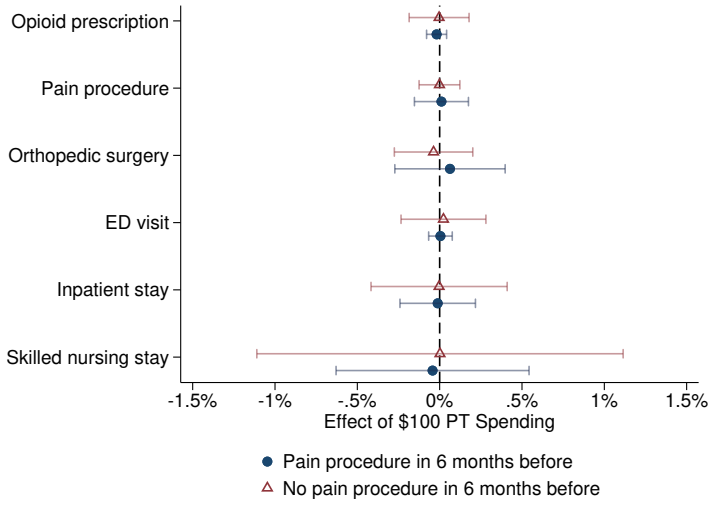
Figure B10: Reduced Form Spending and Health Outcomes



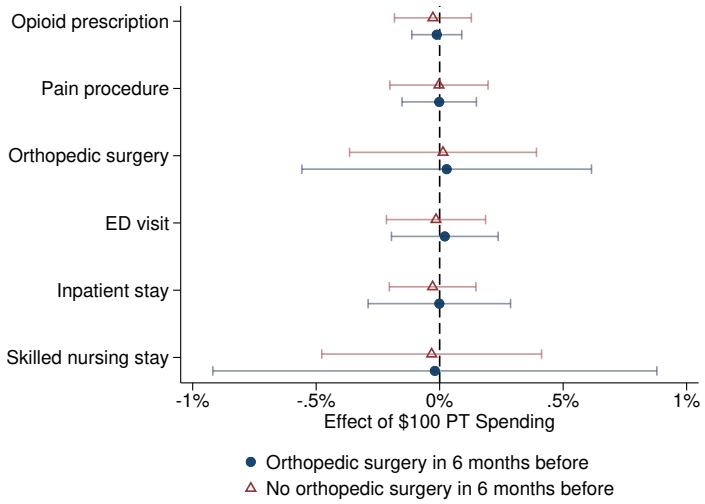
This figure plots the coefficient β, which denotes the effects of an additional \$100 of PT, from estimating Equation 4 on different subsets of predicted need deciles. All outcomes other than opioid prescriptions are measured within 12 months of the first PT session; opioid prescriptions are measured 12-24 months after the first PT session and are scaled in percent terms relative to the sample average. Section F describes how the health outcome measures are defined. Sample is restricted to beneficiaries with less 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



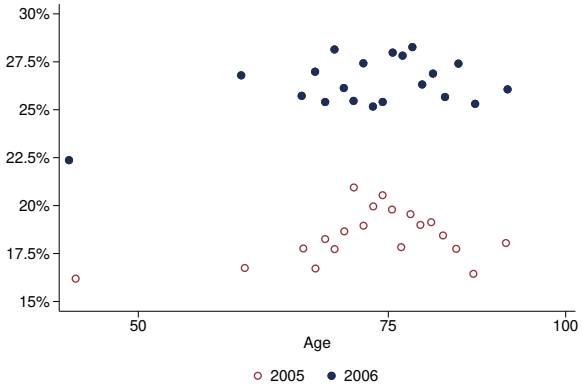
(b) By Recent Orthopedic Surgery



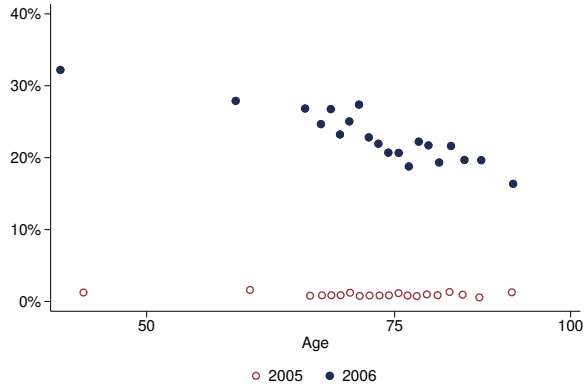
This figure plots the coefficient β, which denotes the effects of an additional $100 of PT, from Equation 4. Results are stratified by whether (a) a patient had a pain procedure in the last 6 months or (b) a patient had an orthopedic surgery in the last 6 months. All outcomes other than opioid prescriptions are measured within 12 months of the first PT session; opioid prescriptions are measured 12-24 months after the first PT session and are scaled in percent terms relative to the sample average. Section F describes how the health outcome measures are defined. Sample is restricted to beneficiaries with less 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: Correlations between patient age and deterrence and denial, 2005-2006

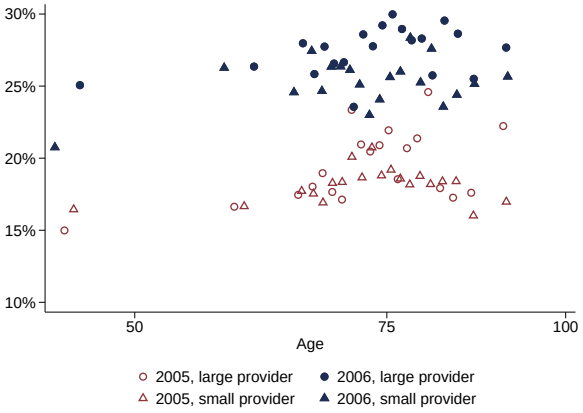
(a) Deterred, conditional on approach



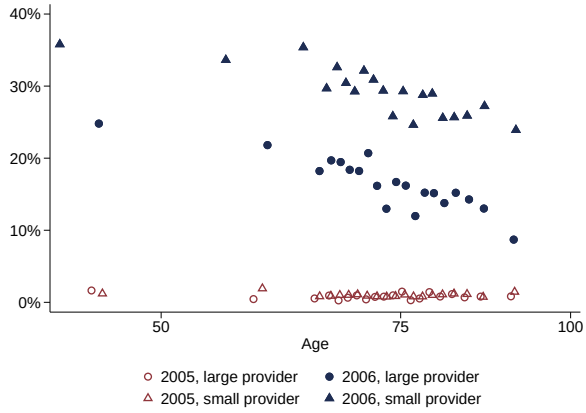
(b) Denied, conditional on attempt



(c) Deterred by provider size, conditional on approach

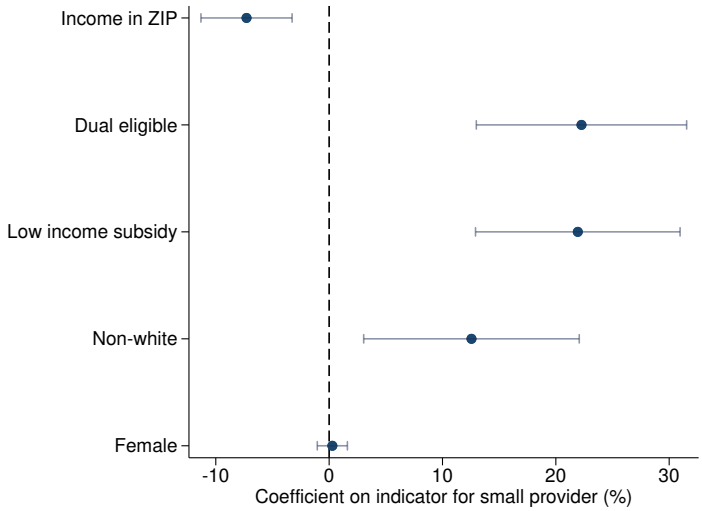


(d) Denied by provider size, conditional on attempt



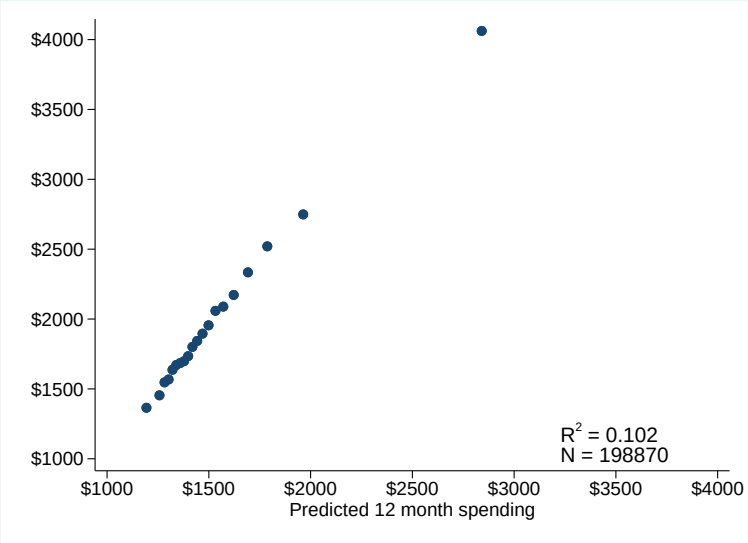
This figure plots the relationship between attempt outcomes and patient need and provider size. "Deterred" is defined as the share of patients who approach the cap but do not attempt, and "denied" is defined as the share of patients who attempt but never make it past the cap. Panels (a) and (b) plot the relationship between log age and share deterred and denied 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.

Figure B13: Correlation Between Provider Size and Patient Demographics



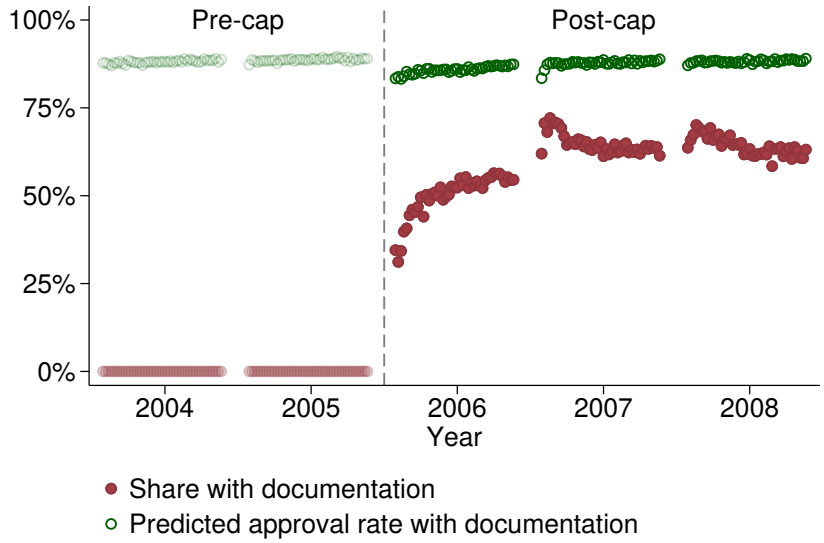
This figure plots the coefficient of the relationship between patient characteristics and an indicator variable for going to a small (below-median size) provider, among patients who approach the cap in 2006 (as defined in Section 2.3). Provider size is measured as a TIN-state's 2006-2008 Medicare patient 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: Predicted vs. Actual 12-month Spending, 2004-2005



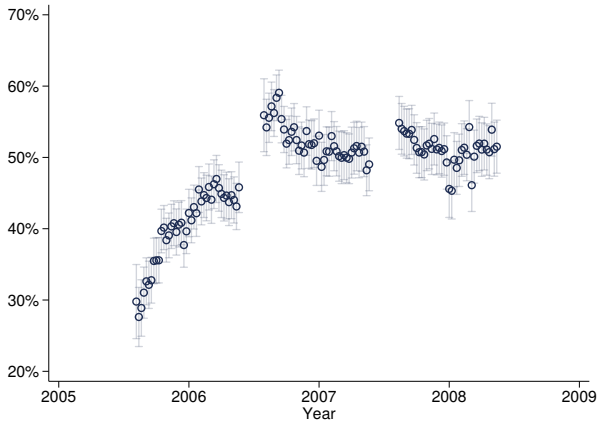
This figure plots the predicted 12-month physical therapy spending against the actual 12-month spending in 2004-2005 from the model described in Section 2.3 and Appendix Section E. Data: 20% Medicare Carrier claims and Master Beneficiary Summary File.

Figure B15: Time Series Variation in Documentation and Predicted Approval Rate on Cap Attempts



This figure plots the share of attempts with at least one claim with documentation as measured by the use of the KX modifier ("Share with documentation") and predicted approval rate based on claim characteristics, assuming documentation ("Predicted approval rate with documentation"). Attempts are defined as weeks in which the patient attempts to go past the cap, where an attempt is as defined in Section 2.3: a week in which a patient's cumulative paid amount is below the cap at the beginning of the week, and their cumulative paid or billed amount is above the cap at the end of the week. The predicted approval rate is derived from the line-level denial rate prediction described in Section E on claims associated with attempts, where the KX modifier indicator for documentation use is set to be on. 20% Medicare Carrier claims.

Figure B16: Industry-Wide Learning Regression Coefficients



This figure plots the unnormalized estimates of industry-wide learning from Equation (7). The 2006-2007 estimates, normalized to 2008, are shown in Figure 10. Data: 20% Medicare Carrier claims.

C Policy Context: Automatic and Manual Process Exceptions

In 2006 only, there was both a "manual" and an "automatic" process for getting an exception for a patient to go past the cap. The manual process required showing documentation and getting pre-approval *prior* to rendering care, while the automatic process meant that providers did not need to get a pre-approval, but could still be subject to *ex post* denial. Patients could qualify for the automatic process by having one of about 100 diagnoses (Centers for Medicare and Medicaid, 2006b). Patients without one of these diagnoses had to go through the manual process, meaning their provider was supposed to wait until the patient was approved before rendering care. Thus, it is possible that some patients who stop at before the cap without an attempt failed to receive pre-approval (which we cannot observe). These patients would be mis-categorized as "deterred" when they were in fact "denied" via the pre-approval process. However, we find that 63% of patients that we classify as deterred have a diagnosis which would qualify them for the automatic exception process, which mitigates this concern. Additionally, the manual exception process states that providers should seek pre-approval "whenever possible" and that "it may sometimes be necessary to request exceptions for services already provided." This suggests that Medicare may have been lenient in enforcing the pre-approval requirement (Centers for Medicare and Medicaid, 2006a).

D Methodology for Assigning Weeks to Cap

In order to characterize provider decisionmaking near the cap, we aggregate the spending to the provider-patient-week-level. This is motivated by evidence that suggests therapy visits are scheduled on recurring days of the week (Figure B1), suggesting that the decision of whether to continue care is made at the weekly level.

We then re-center the panel around the cap so that each observation is in terms of weeks to the cap. This assignment differs depending on whether a patient eventually makes it past the cap. For patients who make it past the cap, we assign week 0 to be the first week they crossed the cap, and assign every other week in which they received PT relative to that.

We cannot directly observe how many weeks a patient is from the cap, so assigning weeks to cap patients who never make it past the cap requires making extrapolation about future weeks of care. We address this by estimating how many additional weeks it would take for a patient to get above the cap based on their prior weekly spending. In particular, we estimate how many weeks it would take based on their own rolling 5-week average or the sample rolling 5-week average, and use the maximum of the two averages. We incorporate the sample rolling average to ensure that patients with idiosyncratically low weekly spending are not assigned to be implausibly far from the cap, and those relatively high average weekly spending are assigned to be closer to the cap.

Additionally, if a provider receives a denial on a bill, the amount that they attempted to bill for is not reported. We address this by imputing the average price that provider was paid for that procedure that year. For patients who attempt to bill over the cap but are denied, the last week that they make an attempt is assigned as week -1, and all other weeks are re-centered relative to that. Patients who never attempt to bill over the cap, we extrapolate based on the last week of care how many additional weeks they would need to reach the cap, and assign all other weeks relative to that.

E Machine Learning Predictions

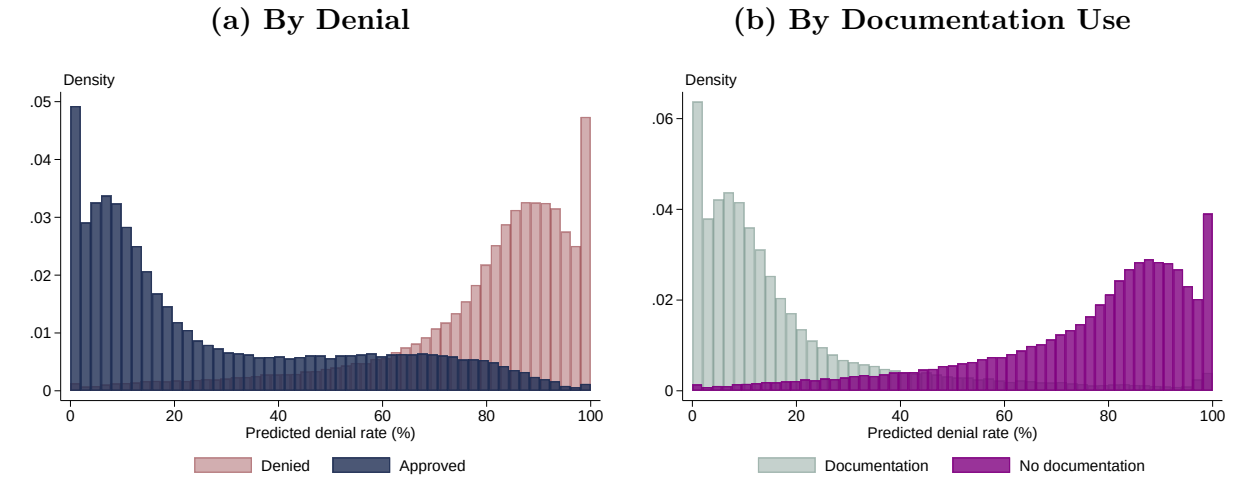
Patient-level 12 Month PT Spending Prediction In order to construct a proxy for patient need, we construct a patient-level measure of predicted PT spending based on patient characteristics and utilization in the 6 months and the year *prior* to starting PT. We then apply this prediction to all 2005-2006 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 or attempt the cap 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, race, 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 B14 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.

Line-level Denial Rate Prediction Table 5 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. So to take this further, we train

a machine learning model that uses line-level information from claims associated with the attempt as well as the weeks of care leading up to the attempt. We first use data from 2006 to train a machine learning model that predicts the likelihood of line-level denial rate for claims aiming to exceed the cap using information from the claim. Specifically, we again use gradient-boosted decision trees from the LightGBM package. The predictors include patient age, race, 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 KX modifier), diagnosis codes, procedure codes, 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. 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 E1: Predictions by Denial and Documentation Status



This figure plots line-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 line was denied or not. Panels (b) plots the modified predicted denial rates after including the documentation modifier code on all lines in the sample, or excluding it. The prediction model is trained on 2006 claims associated with cap attempts and explanatory variables in the prediction model are discussed in Appendix Section E. Section 2.3 describes the definition of cap attempts in further detail. Data: 20% Medicare Carrier claims and Master Beneficiary Summary File.

Figure E1a shows that our model predicts denials out of sample extremely well. For lines

that are denied, the median predicted probability of denial is 79%. We label 90 percent of denied claims as having more than 50% probability of denial and 62% of denied claims as having more than 80% probability of denial. Likewise, for approved claims, the median predicted denial probability is just 16%. We label 78% of these accepted claims as having less than 50% probability of denial, and we label 55% of these accepted claims as having less than 20% probability of denial.

While the model incorporates many more variables than the regression in Table 5, again we find that one factor has outsize predictive influence: documentation. Specifically, whether a claim indicates that documentation of medical need are available to the Medicare contractor reviewing the claim. Figure E1b shows that even though the machine learning model was trained using a host of patient- and claim-level characteristics, this documentation indicator alone explains a substantial fraction of the prediction. Indeed, this is because documentation explains a large amount of variation in the data: in 2006, just 37% of lines involving documentation were denied, while 75% of lines without documentation were denied.

The high degree of predictive power for documentation is consistent with the significant emphasis on documentation in Medicare's public communications about the cap (Centers for Medicare and Medicaid, 2006c). While in many instances, Medicare contractors approved care above the cap solely based on factors easily observable on the claim—such as a patients' diagnoses—documentation gives the provider a means by which to highlight key factors that might have been missed or are difficult to observe from claims alone. For example, a therapist can document that a patient is continuing to make substantive improvements with additional care. Additionally, if it is less costly compile documentation for patients with a sincere medical need, then the mere act of offering documentation can be a credible signal for the contractor reviewing claims.

F 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 MEDPAR and Outpatient files using based on DRG codes and HCPCS codes for surgeries, respectively. Crosswalks to HCPCS codes for pain procedures and PT surgeries and to DRGs for PT 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 nurisng 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).

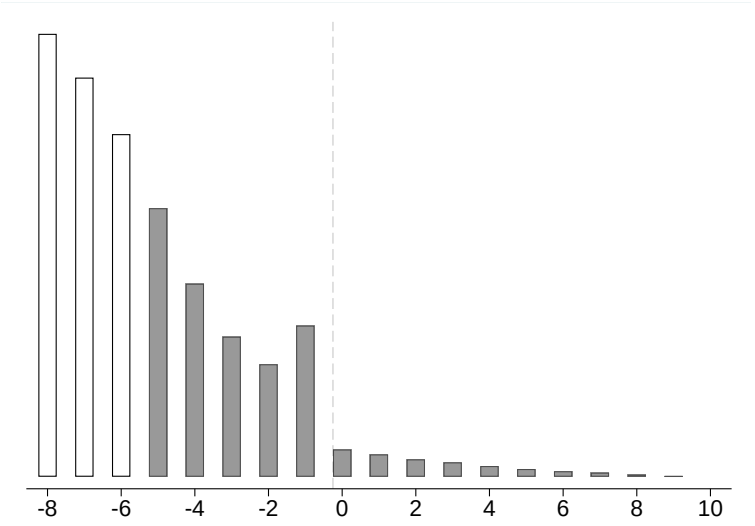
G Patient Health Effects: Difference-in-Difference Strategy

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 also 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 G1 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.[28] The underlying assumption

[28]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

is that the control group does not include any "bunchers" whose spending was at risk of being affected by the cap.

Figure G1: 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 G.

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. This patient contributes to the treated group's average both before and after the reform. Thus, if reducing spending hurts patients, 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 assump-

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.

tions. 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 Appendix 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 whose care in year t ends in calendar week w is:

$$Y_{it} = \beta_0 + \beta_1 Treated_i + \sum_{\tau=2004}^{2008} \beta_{2t}\ Treated_i \times 1(\tau = t) + \lambda_w + \varepsilon_{it}. \tag{9}$$

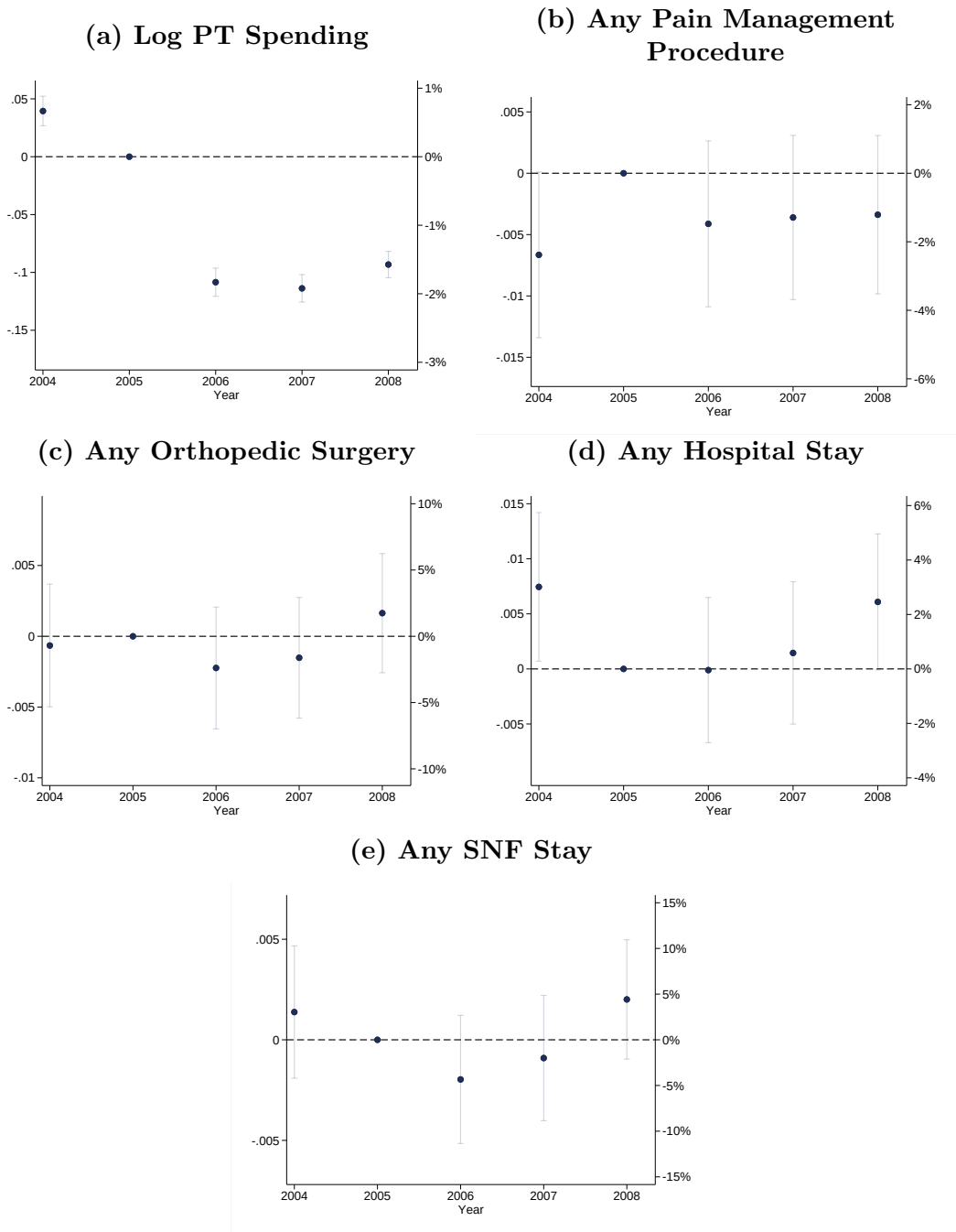
The pooled specification is:

$$Y_{it} = \beta_0 + \beta_1 Treated_i + \beta_2\ Post_t + \lambda_w + \varepsilon_{it}. \tag{10}$$

The estimates are clustered at the patient level and the omitted year is 2005. Figure G2 plots the β_{2t} estimates from Equation 9 and Figure G3 plots the coefficients the pooled specification in Equation 10. Figure G2 panel (a) plots the effects on PT spending and confirms a "first stage" – the treated group saw a sizeable 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.[29]

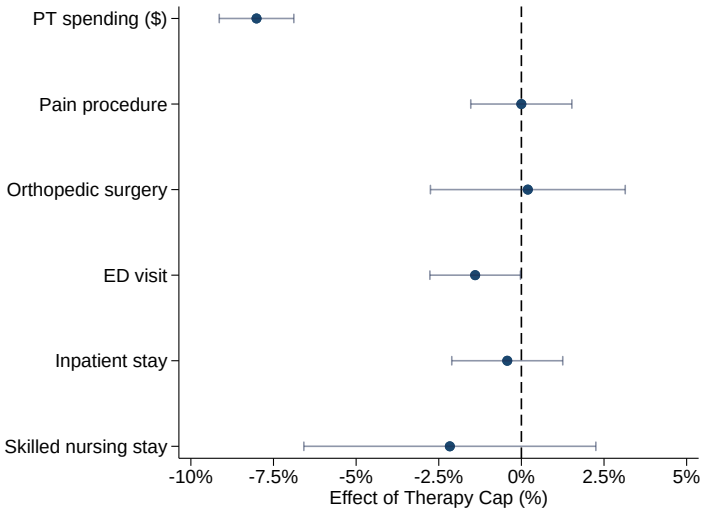
[29]Given that the Medicare Part D program for prescription medication began in 2006, we cannot use this difference-in-difference strategy to study opioid prescriptions.

Figure G2: DD Design: Spending and Health Outcomes within 12 Months of First PT



This figure plots the coefficients from the health outcomes regression in Equation 9. 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 a hospital stay, and panel (e) plots the coefficients on an indicator for a SNF 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 G3: DD Design: Spending and Health Outcomes within 12 Months of First PT



This figure plots the coefficients from the health outcomes regression in Equation 10. 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.