Controlling Costs Through Soft Spending Limits: Evidence from the Medicare Therapy Cap

Maggie Shi and Ashvin Gandhi, Discussed by Justine Mallatt.

October 26, 2024

Disclosure Statement

The views expressed here are those of the authors/discussants and do not represent those of the U.S. Bureau of Economic Analysis or the U.S. Department of Commerce.

- Governments and firms delegate expenditure to agents who may privately gain from spending
- Monitoring every spending decision is infeasible
- One solution: use **soft spending limit** and only scrutinize if spending exceeds limit
  - "Soft" aspect of cap allows exceptions to bypass limit when deemed necessary
  - Examples: Procurement thresholds, project budgets

- Governments and firms delegate expenditure to agents who may privately gain from spending
- Monitoring every spending decision is infeasible
- One solution: use **soft spending limit** and only scrutinize if spending exceeds limit
  - "Soft" aspect of cap allows exceptions to bypass limit when deemed necessary
  - Examples: Procurement thresholds, project budgets
- A way to screen out wasteful spending via monitoring or by inducing efficient sorting around an ordeal
- May give unfair advantage to those who are better at navigating rules

Medicare Soft Cap

Instance of a soft cap in health care:

- Insurers delegate spending to providers, who act on behalf of themselves and patients
- A lot of potential waste to screen out: 25-50% of spending unnecessary (Cutler 2018)
- Small providers are not good at navigating insurer rules (Dunn et al. 2024)

Context: per-patient physical therapy spending cap within Medicare, which can be bypassed if provider can attest to medical necessity.

Medicare Soft Cap

In 2006, Medicare set a soft limit of $1,740 per patient, per year, in PT spending.

$1,740 was about 10 weeks of physical therapy, mostly comprised of therapeutic exercises and manual therapy.

A provider may submit documentation of patient need for PT spending above the cap, which will be reviewed by Medicare.

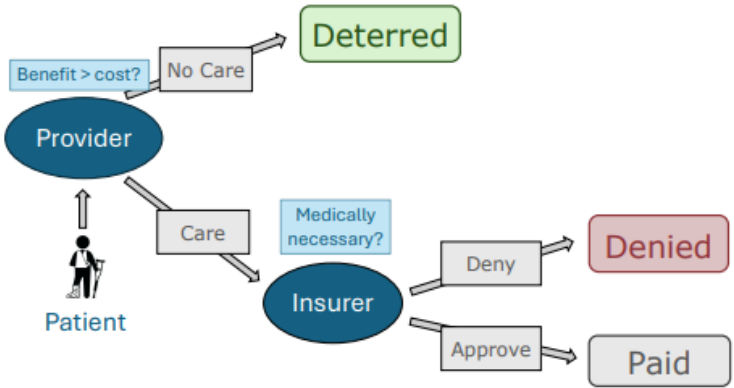


Figure: Flow Diagram of Provider Decision

The authors examine **how** the soft cap on PT spending resulted in Medicare savings. This is a very in-depth investigation into how the policy is working.

- What is the effect of the policy on expenditure?
  - Measure amount of savings.
  - Decompose savings into deterrence and denial channels.
  - Dynamics of savings amount over time

- What impact is the policy having on patients?
  - Policy targeting: does the policy have different effects on sicker vs. healthier patients?
  - Effect on other non-PT health outcomes?

- What impact is the policy having on providers?
  - Heterogeneous effect on providers based on provider size.
  - Large providers better at getting claims approved.
  - Large providers are better at documentation/paperwork.
  - How are firms learning to navigate the documentation process over time?

Medicare 20% sample of claims.

- 2004-2008 to examine soft cap (2006) effect
- 1999-2000 to examine previous hard cap (1999) effect on PT spending

Isolate to beneficiaries with in-office PT.

- At least 5 weeks with $\geq$\$50 a week in the calendar year.
- 186,914 patients in 2006; 169,949 patients in 2005.

Providers are identified by their Tax Identification Number.

- Providers are classified at the practice level.

Work with data at the patient/week level.

- For each patient and PT claim, determine the number of weeks away the patient is from hitting the cumulative annual $1,740 cap when that claim occurs.
- For patients who end year above the cap, week 0 is the week in which their spending crosses the cap.
- For patients who make an attempt but never get past the cap, their last week is week -1.
- For patients who never make an attempt to cross the cap, extrapolate where week 0 is based on past average spending.

An **attempt** to cross the cap threshold when a patient's spending is below the cap at the beginning of the week, and during the week they are *billed* or *paid* such that their cumulative spending is above the threshold at the end of the week.

A patient's **approach** are the weeks before the first attempt (if they attempt) or the weeks before the extrapolated week 0 if they don't end up attempting.

A patient is **denied** if they make an **attempt** to cross the cap but the claim is denied.

A patient is **deterred** if they **approach** the spending cutoff but never **attempt** to cross it.

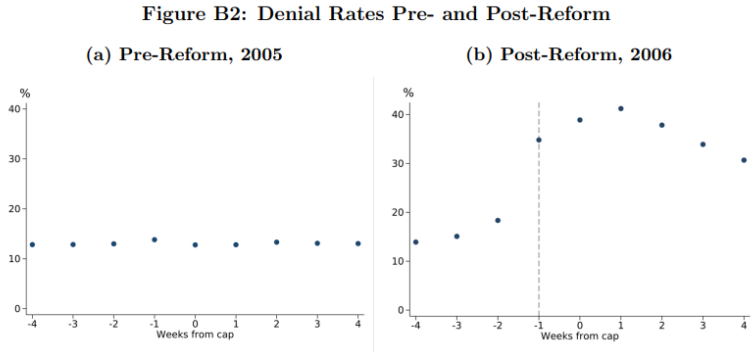


Figure: Denial Rates Around Weeks Around Cap

Effect of Soft Cap on Medicare Spending

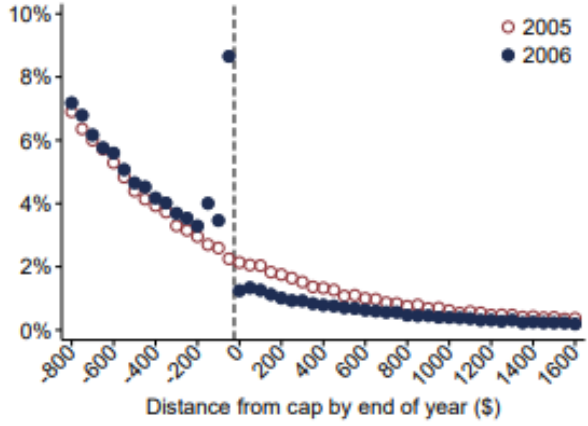


Figure: 2006 Has Bunching in Patient Annual Spending Around the Cap

Placebo – No Difference in 2004 to 2005

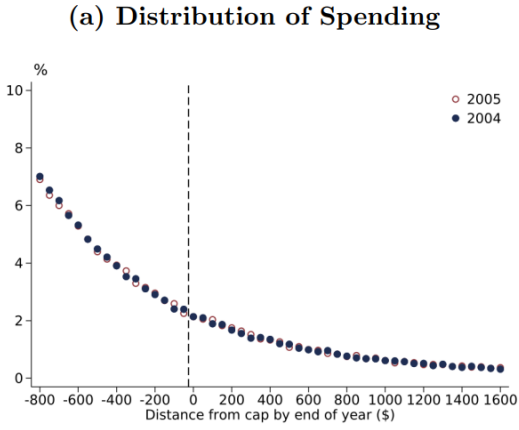
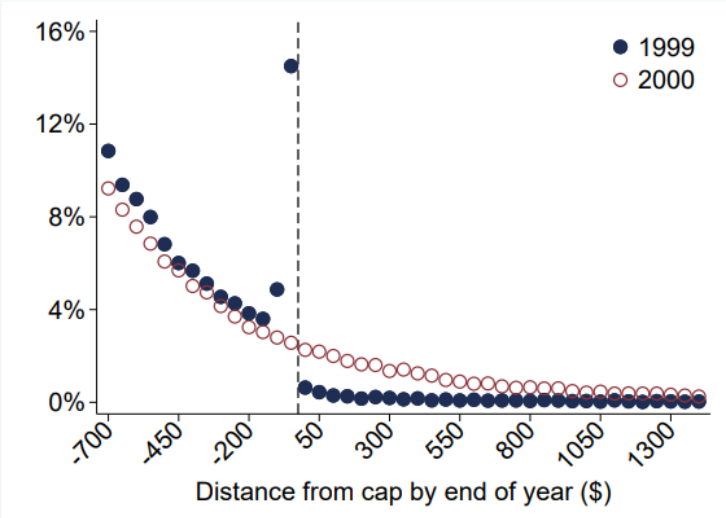


Figure: Bunching in Share of Patients With Spending Around the Cap

Comparison to Hard Cap in 1999



Decomposing into Deterrence vs Denials

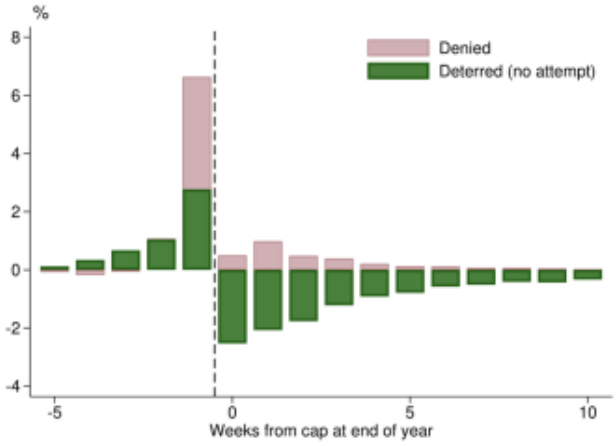


Figure: Difference Between Distributions of Shares of Patients With Spending Around The Cap

The soft cap was meant to curb wasteful spending on PT, while preserving access to treatment for patients with high need.

Are patients that are considered high need (high expected spending) less **deterred** by the policy? Are their claims less frequently **denied**?

Demonstration of Hypothetical Screening by Patient Need

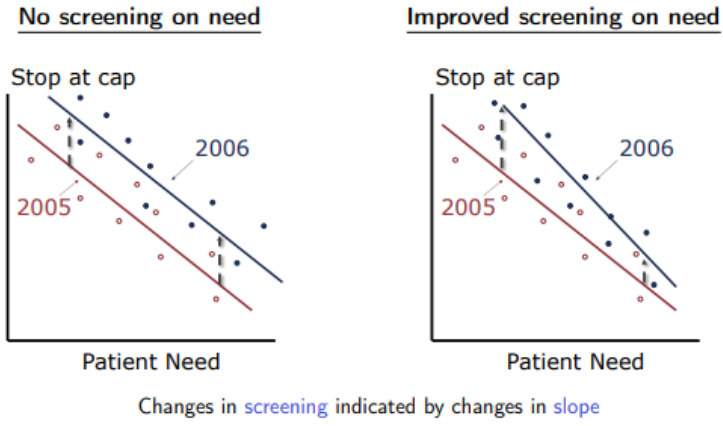
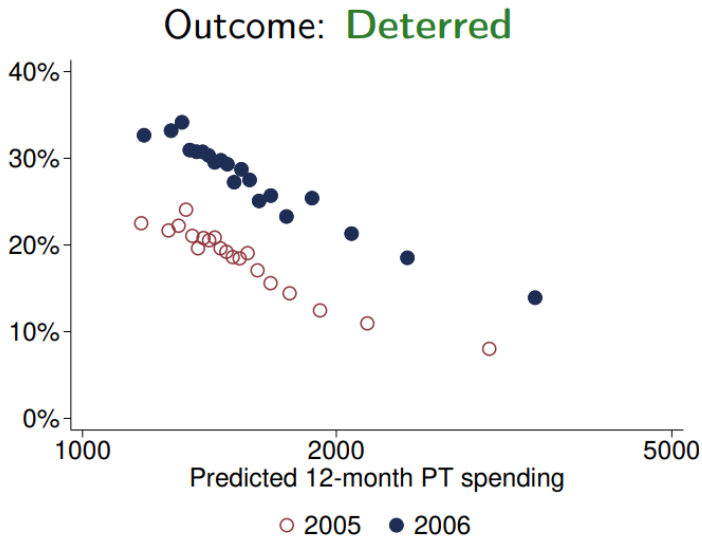


Figure: Changed Slope Shows Targeting Patients with Higher Need

Predict what patient spending would have been in 2006 in the absense of the cap.

Train a ML model on 2004, 2005 to predict what patients in 2006 should be spending. Rank patients based on anticipated need in 2006.

Are Providers/Medicare Screening Based on Patient Need?



Are Providers/Medicare Screening Based on Patient Need?

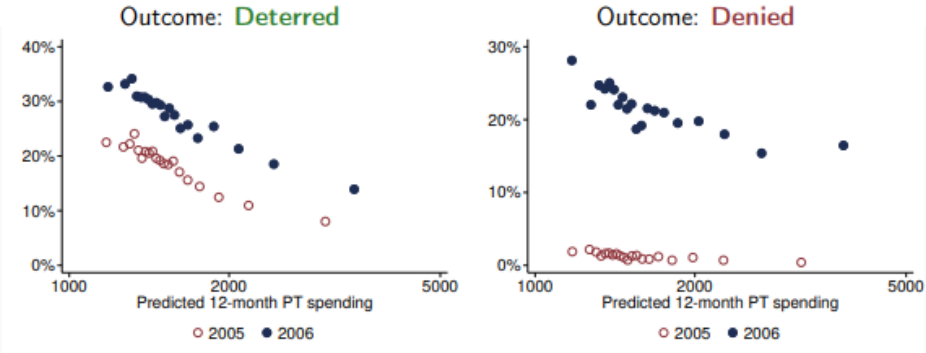
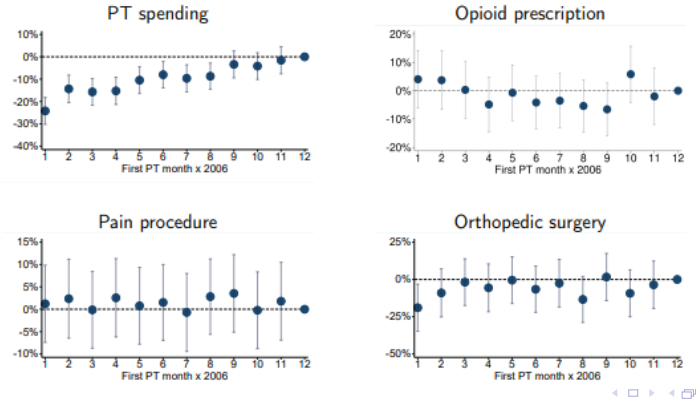


Figure: Targeting in Denials but Not In Deterrence

Difference-in-Differences: Comparing patients that start PT early in the year (and have more time to hit the cap before it resets in January of the following year) to those that start PT later.

First Stage and Reduced Form



- Provider size: How many PT patients does a provider (practice) see?
- The rate of deterrence isnt correlated with provider size, but the rate of denials decreases as provider size increases.

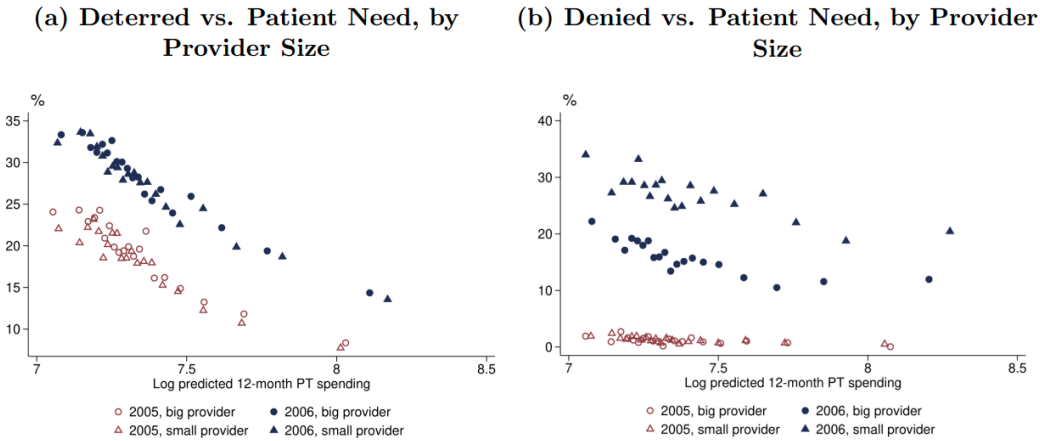


Figure: Provider Size and Patient Need

Provider Size and Documentation

- Large providers less frequently denied.
- Authors argue that large providers are better at **documentation**.
- Documentation measurement: Providers must use a "KX" modifier code to attest to medical necessity of spending above limit.

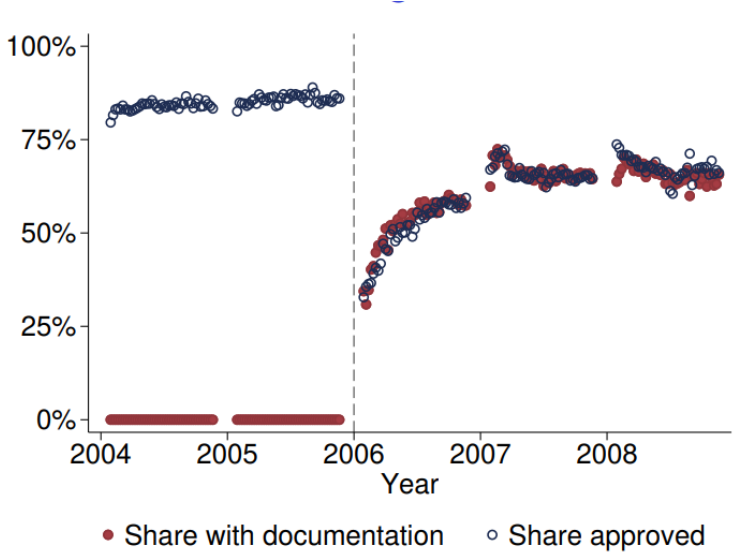
Documentation Matters

- In a regression on denial probability, adding documentation increases the model fit most.
- ML model predicting denials suggests that documentation on a claim is most predictive of denial/approval.

	(1)	(2)	(3)	(4)
		Approved		
Has documentation	0.296*** (0.00438)	0.290*** (0.00449)	0.296*** (0.00443)	0.281*** (0.00445)
N	166808	142402	166809	142402
Patient demographics	X	X		X
Patient health		X		X
Provider size			X	X
Week FE	X	X	X	X
R^2 including documentation	.212	.21	.209	.226
R^2 *excluding* documentation	.039	.042	.036	.07

- In contrast, patient characteristics have almost no power in measuring denials.

Documentation and Denial Time Series



Fit a linear model at the patient/provider level explaining the probability of denial/approval based on:

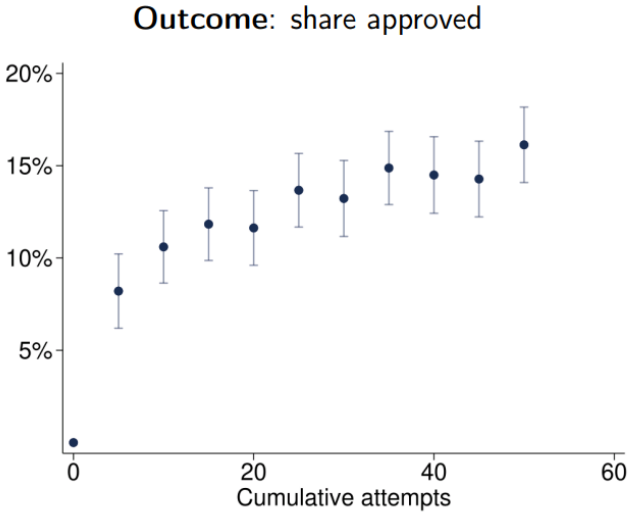
- Time invariant patient characteristics
- Patient medical consumption patterns that vary over time.
- Provider FE (size enters here)
- Time FE to capture industry-wide learning about documentation
- **Experience: Each providers' cumulative number of attempts to cross soft cap**

- **Approach**: regression of dynamics of approval dynamics for patient i, provider j, in week t.

$$Approval_i = \underbrace{\beta_{N_{jt}}}_{\substack{\text{Cumulative attempts:} \\ \textbf{experience}}} + \underbrace{\alpha_j}_{\substack{\text{Provider FE:} \\ \textbf{fixed advantage}}} + \underbrace{\gamma_t}_{\text{Time FE}} + X_i + \epsilon_i$$

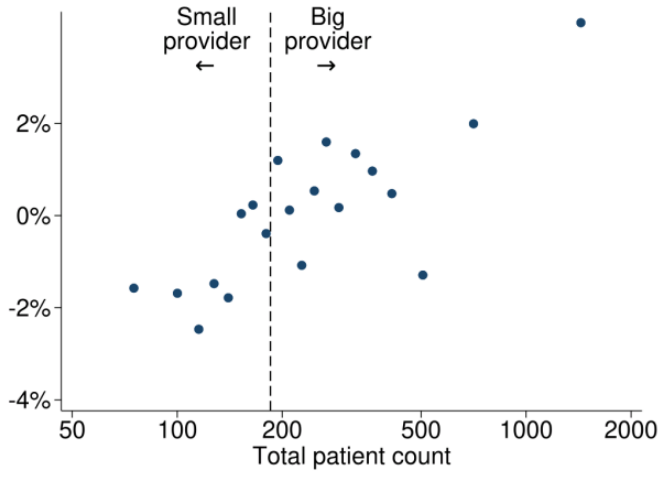
- **Sample**: all attempts to bill past cap in 2006-2008

Firms Learn with Experience

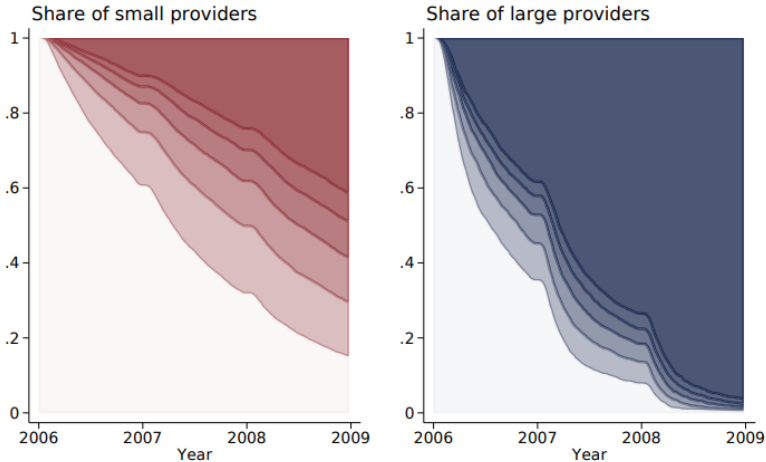


Large Firms Have a Time Invariant Advantage

Outcome: provider FE on share approved



(a) Experience over Time



What happens when a provider successfully reverses a denial by adding documentation?

Event study around likely "learning event": first time provider corrects a denial

Outcome: share using documentation

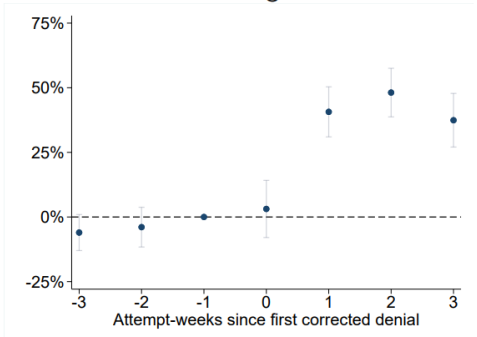


Figure: Evidence of A Learning Event

As providers get better at documentation, denial rates fall over time which reduces the savings from the soft cap.

Figure 5: Effect of the Cap from 2006 through 2008

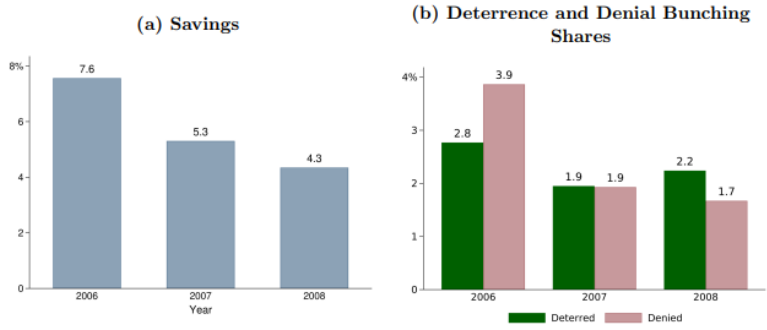


Figure: Estimated Savings from Bunching Models, 2006-2008

Takeaways: Spending

The soft spending cap on PT reduces spending for patients.

- 8% savings; $152 per additional denial.
- Deterrence accounts for 40 percent of savings; denials for 60 percent.

Evidence that providers are deterred.

Providers are less likely to attempt to cross the soft cap of spending with patients of all need levels; there's not evidence that providers are discriminating based on patient need.

Medicare is more likely to approve attempts from patients that are measured as high need.

- Policy screening burden is on Medicare, not on providers.

Takeaways: Documentation and Large Providers

Large providers have an advantage in successfully documenting patient need.

- Large providers have a fixed advantage in getting claims past the cap approved.
- Larger providers accumulate experience with claims around the spending cap more quickly.

The soft cap has created savings for Medicare and:

- Doesn't lead to spillovers in other patient treatment options.

The soft cap has created savings for Medicare but:

- Has created compliance costs for providers that has deterred both healthier and less healthy patients from receiving care over the cap.
- Placed a cost of screening onto Medicare's adjudication process.
- Unintentionally favors large providers.
  - The policy "introduced horizontal inequity along a dimension unrelated to patient need."
- Had diminishing effectiveness in creating savings over time as providers learned to navigate paperwork.

My suggestions are minor, my job is easy, the paper is excellent.

Amazing paper!

This is some of the best applied work on healthcare claims I've ever seen.

The paper is extremely clear, which is impressive because this paper:

- Demonstrates deep knowledge of claims data.
- Reveals meticulous consideration of the policy and how it works.
- Takes full advantage of a broad range of econometric methods (regression, diff-diff, bunching models, machine learning, and more)
- Is extremely well-written. Clear on methods, frames work within the literature well.
- Presents results in visually compelling ways.

This is a sophisticated paper that is clear and polished.

Recommendation: Fleshing out Deterrence Channel

Deterrence falls between 2006 and 2007.

The paper shows that providers of all sizes are deterred from even attempting care over the soft cap. Moreover, providers are bad at deciding which patients they'll attempt with. Is there evidence that providers thought this was a hard cap?

- Evidence of learning?
- E.g. Use a sharp learning event to see if a provider's pool of claims bunching up to the cap suddenly changes. More bunching before vs after the sharp learning event? Look at if the ratio of attempts to approaches changes suddenly?
  - If there is learning around deterrence, does that impact indiscriminate targeting across patient need? Does it interact with provider size?

.

Recommendation: Checking for Changes in "Upcoding"

Providers may think they need to offer "proof" of medical necessity beyond adding the "KX" documentation. Do providers tend to bill for more severe diagnosis codes as patients approach the soft cap?

Alternatively, providers may think they need to tread lightly around the cap, and may use less severe diagnosis codes thinking smaller billed amounts will be more likely to be approved.

If providers are billing more/less aggressively to surpass the cap, this behavior may impact Medicare savings.

Bunching model holds prices constant by using average billing/paid amounts in claims by procedure and diagnosis codes? Authors may be shutting down part of the story with this method.

Minor Suggestions

- A few more sentences explaining the intuition of the ML model in the appendix.
- Figure 7b appears to have one part of it where 2004, 2005 are given a theoretical denial/approval probability based on a model trained on 2006? Maybe drop this, the documentation story is convincing without it.
- Appendix C mentions a set of 100 diagnosis codes that get claims on the fast track to being automatically approved – does this interfere with results?

Thank you!

Additional questions and comments?

Thank you for having me!

Justine Mallatt, Research Economist at Bureau of Economic Analysis.

Justine.Mallatt@bea.gov