

Redesigning Graduate Medical Education Payments: The Role of Residency Caps

Thomas Yu*

October 20, 2026

[Click here for latest version](#)

Abstract

Physicians in the U.S. must complete graduate medical education (GME) by training as residents in teaching hospitals. Teaching hospitals receive GME payments from Medicare, with each hospital assigned a cap on its subsidized residents. About half of teaching hospitals exceed their caps and thus receive no additional payments for training more residents. I study whether GME payments can be redesigned to increase the number of residents by redirecting payments from over-cap to under-cap hospitals. Instrumenting for under-cap status using changes in hospital size since the caps were assigned, I find that under-cap hospitals' resident counts respond to payments with an elasticity of 0.74, while over-cap hospitals have no response. My estimates suggest that reducing over-cap hospitals' payments by 25% and transferring the savings to under-cap hospitals would lead to 3,700 more residents without any change in government spending. These gains are proportionally larger in rural areas, which face more severe physician shortages.

*Yale University, t.yu@yale.edu

1 Introduction

Before physicians can practice medicine independently in the United States, they must complete graduate medical education (GME) by training as residents in teaching hospitals. Teaching hospitals receive GME payments through the Medicare program. These payments represent a substantial government expenditure: in 2024, 1,355 teaching hospitals trained 136,824 residents and received \$22.5 billion in GME payments.

GME payments are especially important in light of physician shortages. Total physician supply is projected to fall short of demand by 12% by 2038, with larger shortfalls in rural areas (58% of demand unmet) than in urban areas (5% of demand unmet) (HRSA, 2025). Residency positions represent a bottleneck in the production of physicians, and GME payments are the main federal policy lever to affect those positions. In principle, payments can mitigate physician shortages by encouraging hospitals, especially rural hospitals, to train residents. However, the current design of payments may not do so cost-effectively.

GME payments scale with hospitals' resident counts, with each hospital assigned a cap on its subsidized residents. Most hospitals' caps are based on their 1996 resident counts. As of 2024, 47% of teaching hospitals train residents in excess of their caps, so they receive no additional payments for training more residents. In other words, their marginal payments are zero, even if their total payments are not. This raises the question of whether, for the purpose of encouraging resident training, it would be more cost-effective to reallocate their payments. In this paper, I study whether GME payments can be redesigned to increase the number of residents without any change in government spending by redirecting payments from over-cap to under-cap hospitals.

I begin by presenting a stylized model of the market for residents. In the model, GME payments increase the marginal value of residents, but only for hospitals that receive marginal payments. The model's key prediction is that payments to under-cap hospitals increase their resident counts, with greater per-resident amounts leading to larger increases, but payments to over-cap hospitals have no effect. Thus, redirecting payments from over-cap to under-cap hospitals can increase the total number of residents.

Then, I test and quantify the stylized model's prediction. Using hospital cost report data from the Centers for Medicare & Medicaid Services, I estimate how resident counts respond to GME payments. My sample consists of hospital-years in 2011–2024 for hospitals that have been teaching since 1996, so that their caps are based on their 1996 resident counts. I construct each hospital-year's *effective per-resident amount*, the additional GME payments that the hospital would receive for one more resident if it were under its cap. The effective per-resident amount combines two GME payment programs. A hospital's direct payments scale with its residents,

policy-assigned per-resident amount, and Medicare share. A hospital's indirect payments scale with its resident-to-bed ratio and total Medicare discharge payments. Both components are essential to capturing a hospital's full exposure to payments: indirect payments account for 68% of the average effective per-resident amount in 2024.

My first identifying assumption is that conditional on observed characteristics, effective per-resident amount is correlated with resident count only through the GME payment channel. This is plausible because most hospitals' policy-assigned per-resident amounts are based on historical costs and idiosyncratic accounting practices in 1984, and there is no obvious reason for a hospital's exposure to Medicare patients to affect its resident count, conditional on hospital size and character, other than through GME payments.

Under-cap status is endogenous because it is directly linked to resident count. To address this, I construct instruments using changes in hospital size since 1996. I show that the instruments are relevant: hospitals with more growth since 1996 are less likely to be under their caps. My second identifying assumption is that conditional on observed characteristics, including present-day hospital size, the instruments are correlated with resident count only through the under-cap status channel. Intuitively, in a comparison between two hospitals with identical present-day characteristics but only one of which grew since 1996, the only reason for their resident counts to differ is that the growing hospital was assigned a lower cap and thus is less likely to be under it.

To support my two identifying assumptions, I perform placebo tests using outcomes related to hospital workers excluding residents. I find no evidence that either effective per-resident amount or any of the constructed instruments is correlated with non-resident hospital workers or labor hours, conditional on observed characteristics.

My main results are consistent with the stylized model's prediction. Under-cap hospitals' resident counts respond positively to their effective per-resident amounts with a statistically significant estimated elasticity of 0.74. For over-cap hospitals, the estimated elasticity is negative and I cannot reject that it is zero. I interpret my estimates as elasticities of resident demand with respect to GME payments. This interpretation relies on the presence of excess supply in the market for residents: in 2024, almost all first-year residency positions in the National Resident Matching Program were filled while a notable number of active residency applicants were left unmatched.

Finally, I use my estimates to conduct policy exercises, using the market for residents in 2024 as a baseline. First, as an unrealistic benchmark, I consider removing all caps. A rough linear extrapolation of my estimates suggests that this would increase total resident demand by about 65,000, by proportionally more in urban areas than in rural areas. If this additional demand were filled, total GME payments would increase by \$17.3 billion. However, the number

of unmatched residency applicants suggests that not all of it would be filled.

Then, I consider a more realistic budget-neutral policy: reducing all over-cap hospitals' GME payment rates and using the savings to raise all under-cap hospitals' effective per-resident amounts. A 25% reduction for over-cap hospitals would transfer \$3.3 billion, leave total government spending unchanged, and lead to about 3,700 more residents. Because under-cap hospitals are more likely to be rural, these gains are proportionally larger in rural areas: 5.3% in rural areas and 2.7% in urban areas. The proposed redesign creates opportunities for residents in areas facing more severe physician shortages.

Related literature. First, I contribute to the literature studying the effects of GME payments. Newhouse and Wilensky (2001) review the economic justification for subsidizing teaching hospitals and the debate over the design of Medicare's role. Nicholson and Song (2001) show that the introduction of indirect GME payments led hospitals to train more residents without changing Medicare admissions or use of nurses.

More recent work has focused on the role of residency caps, and in particular on policy-driven redistributions of cap slots. McNamara and Hussain (2025) find that a hospital awarded one additional slot in 2010 expanded by roughly one resident, so hospitals near their caps take up additional capacity close to one for one. McNamara and Pineda-Torres (2025) find that new cap slots draw residents into primary care and raise the supply of primary care physicians in affected counties, evidence that expansions of subsidized training translate into practicing physicians where they occur. These papers identify the effect of relaxing the cap for hospitals that were at or near it. In this paper, I ask a complementary question: taking caps as given, how does a hospital's response to its effective per-resident amount vary depending on its position relative to its cap? My estimates suggest that the hospitals that respond to payments are precisely those with room below their caps. This variation is the key to designing a more cost-effective reallocation of payments.

A parallel policy literature has proposed making GME payments more effective by tying them to performance or workforce needs (Institute of Medicine, 2014). My proposed redesign is simpler and budget-neutral: it changes only how existing payments are divided between hospitals that do and do not receive marginal payments.

Second, I contribute to the literature on the resident and physician labor markets. Gottlieb and Nicholson (2026) describe how entry into medicine is rationed by caps on medical school seats and on residency slots, while trained physicians increasingly work alongside nurse practitioners and physician assistants, whose numbers have grown far faster. Nicholson (2003) shows that barriers to entering the better-paid specialties are large enough that medical students would be willing to pay for those residency positions, which implies an excess supply of

applicants and is what allows me to read my estimates as demand elasticities. In terms of policy interventions, Gottlieb et al. (2025) show that payments to physicians shape their earnings, labor supply, and specialty choices, so the number of positions a payment reform creates is not the end of its effects. My contribution is to propose a novel policy that targets a specific bottleneck, residency positions, and to show that it can be implemented in a budget-neutral way that disproportionately affects areas facing more severe physician shortages.

Third, I contribute to the literature studying how labor demand responds to financial incentives and whether or not those incentives operate at the margin. Saez et al. (2019) find employment gains from a payroll tax cut for young workers in Sweden, a marginal subsidy. Cahuc et al. (2019) study a French hiring credit paid only on new hires during the Great Recession, find sizable employment effects, and show that its effectiveness depended on being targeted at hires rather than at the existing workforce. Where subsidies are instead paid for hires that would have occurred anyway, the effects are much weaker: Jain et al. (2026) study the Work Opportunity Tax Credit in the U.S. and find precise null effects on hiring, suggesting that the marginal incentive of the credit is zero even if the total incentive is not.

In many settings, identifying marginal incentives is difficult; firms that face marginal incentives often differ from those that do not for reasons tied to the outcome. GME payments provide a well-suited case study: the same payment formula applies to every teaching hospital, and marginal payments are determined by a threshold fixed by each hospital's 1996 resident count. In this paper, I provide an estimate of how much of an incentive's effect is lost when it no longer operates at the margin, and to what extent it can be recovered by reallocating it toward recipients exposed to marginal effects.

Outline. In Section 2, I go over background information related to GME payments, the data that I use, and descriptive facts about the market for residents. In Section 3, I present a stylized model that illustrates the relationship between caps and the effectiveness of payments. In Section 4, I outline my empirical strategy and results. In Section 5, I conduct policy evaluations and propose a redesign that would increase the number of residents without any change in government spending. In Section 6, I conclude.

2 Background, Data, and Descriptive Facts

2.1 Background

Residency training. After completing medical school, physicians must complete graduate medical education (GME) before practicing medicine independently. GME consists of training

as a resident in a specialty, which can last from 3 years in primary care fields to 7 years in some surgical fields, followed by the option of a fellowship in a subspecialty, which can last from 1 to 3 years, all of which takes place in teaching hospitals. Most first-year positions are filled through the National Resident Matching Program, a centralized clearinghouse that matches applicants to programs based on both sides' rank order lists. In 2024, 44,853 active residency applicants submitted rank order lists for 38,494 first-year residency positions; 99.6% of positions were filled while 6,678 active applicants were left without a first-year position, suggesting the presence of excess supply in the market for residents (NRMP, 2024). In this paper, I follow the standard established by Medicare's payment rules of referring to all GME trainees, including fellows, as "residents."

Why GME is subsidized. Residents provide patient care and receive wages while they train. Training is general (i.e., valuable to residents even if they switch hospitals), and it is costly for hospitals to provide because it requires faculty supervision, administrative infrastructure, and slower and more resource-intensive care. In the absence of frictions, residents would bear these costs through lower wages, and the market for residents would clear. In practice, this may not happen for many reasons; for example, institutional norms encourage hospitals to pay a minimum reasonable wage and keep wages equal across very different specialties, and residents face borrowing constraints after medical school (Nicholson, 2003). As a result, hospitals' demand for residents may be lower than the welfare-maximizing level, and a subsidy can raise the number of residents toward the efficient outcome. Medicare's GME payments to teaching hospitals are the largest such subsidy and the focus of this paper.¹

How payments are calculated. Medicare pays teaching hospitals through two programs. Direct graduate medical education (DGME) payments are intended to offset the direct costs of training, such as residents' salaries and supervising physicians' time. In general, hospital i 's DGME payments in year t are

$$DGME_{it} = PRA_{it} \times McrShare_{it} \times \min(Residents_{it}, Cap_i^{DGME}) \quad (1)$$

where PRA_{it} is the hospital's policy-assigned per-resident amount and $McrShare_{it}$ is the share of its inpatient days attributable to Medicare patients.

Most hospitals' policy-assigned per-resident amounts are computed by dividing accounting costs that were recorded as GME-related in the hospital's 1984 cost report by its resident count in that period, updated for inflation. Primary care residents receive a slightly higher per-resident

¹Teaching hospitals also receive other subsidies, such as GME payments from state Medicaid programs and, for children's hospitals, the Children's Hospitals GME program. These are notable but outside the scope of this paper.

amount than other residents. Because these amounts reflect idiosyncratic accounting decisions made four decades ago, two hospitals training identical residents today can be assigned very different amounts (Newhouse & Wilensky, 2001).

Indirect medical education (IME) payments are intended to compensate teaching hospitals for the higher patient care costs associated with training. They are a percentage add-on to Medicare's payment for each inpatient discharge. In general, hospital i 's IME payments in year t are

$$\text{IME}_{it} = c_t \times \text{McrDischarge}_{it} \times \left[\left(1 + \frac{\min(\text{Residents}_{it}, \text{Cap}_i^{\text{IME}})}{\text{Beds}_{it}} \right)^{0.405} - 1 \right] \quad (2)$$

where c_t is a multiplier set by statute (1.35 since 2008), McrDischarge_{it} is the hospital's total Medicare discharge payments, and Beds_{it} is the hospital's bed count. Of the two programs, IME payments are larger: in 2024, teaching hospitals received \$16.2 billion in IME payments and \$6.3 billion in DGME payments.

Caps. The Balanced Budget Act of 1997 limited the number of residents for which each hospital is paid to its resident count in 1996. Because DGME and IME payments count residents slightly differently (e.g., different weights on residents and fellows), they have separate caps, $\text{Cap}_i^{\text{DGME}}$ and $\text{Cap}_i^{\text{IME}}$. Caps have a stark implication for marginal payments: a hospital above its caps receives nothing for training one more resident.

While caps have remained largely unchanged since their establishment, there have been a few policy adjustments at the margin. Section 422 of the Medicare Modernization Act in 2003 and the Affordable Care Act in 2010 redistributed unused slots from hospitals below their caps to other hospitals, and the Consolidated Appropriations Act in 2021 created new slots. Slots received under Section 422 are paid at lower rates than the rest of the cap. Furthermore, hospitals that begin training residents after 1996 are assigned caps based on their resident counts after a five-year start-up period.

Additional features. Three technical details of GME payments matter for measurement. First, resident counts are measured in units of full-time equivalents (FTEs), such that a resident trained part-time or for only part of the year counts as less than 1 resident. Second, the capped resident count that enters each formula is actually a three-year rolling average of capped resident counts. Third, for DGME, resident counts are weighted (a fellow counts as half of a resident). In this paper, I use unweighted resident counts as the measure of residents.

2.2 Data

My main data come from annual hospital cost reports provided by the Centers for Medicare & Medicaid Services through the Healthcare Cost Report Information System. I use data in 2011–2024, from the first year of the current cost report format to the last year with complete data. After extracting resident counts, caps, policy-assigned per-resident amounts, Medicare shares, Medicare discharge payments, and beds, I reconstruct each hospital’s GME payments by applying the statutory formulas, including the rolling average calculation, the differential treatment of fellows and primary care residents, and the Section 422 slots. All monetary variables are adjusted for inflation using the consumer price index and expressed in 2024 dollars. Appendix A describes the construction and validation of these variables in detail.

A hospital is considered a teaching hospital if it has a positive resident count. In 2024, 1,355 teaching hospitals trained 136,824 residents and received \$22.5 billion in GME payments. For estimation, I restrict to hospitals that have been teaching since 1996 because their caps are based on their 1996 resident counts, a key premise of my estimation strategy. The 1996 resident counts come from the data released under the Consolidated Appropriations Act in 2021.

I construct each hospital-year’s *effective per-resident amount*, the additional payment the hospital would receive for one more resident if it were under both of its caps, ignoring the rolling average calculation and holding fixed its Medicare share, Medicare discharge payments, beds, share of primary care residents, and share of fellows. In general, hospital i ’s effective per-resident amount in year t is

$$\text{EffPRA}_{it} = \underbrace{\text{PRA}_{it} \times \text{McrShare}_{it}}_{\text{DGME payments}} + \underbrace{\frac{c_t \times \text{McrDischarge}_{it} \times 0.405 \times (1 + \bar{r})^{-0.595}}{\text{Beds}_{it}}}_{\text{IME payments}} \quad (3)$$

Because the IME payment formula is nonlinear in the resident-to-bed ratio, evaluating a hospital’s effective per-resident amount at its own ratio would make it a function of resident count. Instead, I evaluate it at the average resident-to-bed ratio across all hospital-years, $\bar{r} = 0.28$. In 2024, the average effective per-resident amount is \$182,998, and IME payments account for 68% of that average.

A hospital is considered under-cap if either of its resident counts is below its corresponding cap, including bonus cap slots from Section 422. In other words, a hospital is under-cap if it receives any marginal payments, even if they may not equal its full effective per-resident amount as defined in Equation 3, and a hospital is over-cap if its marginal payments are zero. A new teaching hospital, in the five years before its cap is set, earns its full effective per-resident amount on every resident and is considered under-cap.

Finally, I use additional observed characteristics. From the cost reports, I observe average hospital wages, non-resident hospital workers, and non-resident labor hours. From the American Hospital Association Annual Survey, I observe counties, ownership (i.e., non-profit, for-profit, or public), medical school affiliation status, bed counts (which I use as measures of hospital size), and registered nurses. I classify counties as urban (metropolitan) or rural (not metropolitan) according to the Office of Management and Budget's delineation.

2.3 Descriptive Facts

Table 1 summarizes teaching hospitals in 2011–2024: 16,783 hospital-years at 1,566 hospitals. 56% of all hospital-years in that period are over-cap by my definition. My estimation sample restricts to hospitals that have been teaching since 1996. These hospitals are more likely to exceed their caps: 62% of hospital-years in my estimation sample are over-cap.

Over-cap hospitals are larger (as measured by beds) and train more residents than under-cap hospitals, but they have similar effective per-resident amounts. This similarity reflects the fact that effective per-resident amounts are driven by historical accounting practices and exposure to Medicare patients rather than by size or cap status. Under-cap hospitals have shrunk in size since 1996, as evidenced by their change in logged beds, while over-cap hospitals have grown. Conditional on present-day size, this difference in growth is the basis of my instruments.

Figure 1 shows how the share of teaching hospitals over their caps has evolved. The share has drifted down over time, from 62% to 47% between 2011 and 2024. Hospitals that were teaching in 1996 are much more likely to be over-cap than those that began teaching later, whose caps were set more recently and which are often still growing into them.

Figure 2 shows the locations of teaching hospitals by cap status in 2024. Over-cap and under-cap hospitals are found in every region, and in most metropolitan areas side by side, so cap status is not a proxy for location. The share of hospitals over-cap ranges from 41% in the Northeast to 55% in the West, with 51% in the Midwest and 48% in the South. The clearer geographic difference is between urban and rural hospitals: in 2024, 8.8% of under-cap hospitals are in rural areas, as opposed to 4.9% of over-cap hospitals.

Additional descriptive facts are shown in the appendix. Table B1 shows how the marginal payments hospitals actually receive compare with their effective per-resident amounts. Among teaching hospitals that receive GME payments, those below all caps receive their full effective per-resident amount, and those below some caps receive most of it. Those above all caps receive nothing. Figure B1 shows the distribution of DGME residents relative to DGME caps in 2024. The distribution is spread widely on both sides of the cap, with a large mass just above it. There is a long tail of hospitals that far exceed their caps.

Table 1: Summary Statistics

	All Teaching Hospitals			Sample
	All	Under-cap	Over-cap	All
Hospital-years	16,783	7,067	9,419	10,037
Hospitals	1,566	1,222	1,161	784
<i>Mean Per Hospital-year</i>				
Residents	97.6	57.1	130.9	124.6
GME Payments (\$1m)	14.7	11.1	17.9	19.8
DGME (\$1m)	4.4	3.6	5.2	5.9
IME (\$1m)	10.3	7.4	12.8	14.0
Effective PRA (\$1k)	171.3	174.8	168.7	187.3
DGME (\$1k)	57.3	61.2	54.3	60.7
IME (\$1k)	114.0	113.6	114.3	126.6
Beds	364	309	410	412
$\Delta \ln(\text{Beds})$ Since 1996	0.02	-0.06	0.08	0.00
Non-profit	0.72	0.68	0.74	0.75
For-profit	0.13	0.18	0.10	0.10
Med School Affiliated	0.80	0.76	0.85	0.87
Rural	0.05	0.08	0.04	0.04
<i>Total Per Year</i>				
Residents (1k)	117.0	28.8	88.1	89.4
GME Payments (\$1b)	17.7	5.6	12.1	14.2
DGME (\$1b)	5.3	1.8	3.5	4.2
IME (\$1b)	12.4	3.8	8.6	10.0

The first three columns cover 2011–2024 for all teaching hospitals. A hospital is under-cap if it receives any marginal payments. The under-cap and over-cap columns exclude 297 hospital-years in which the hospital trained residents but received no GME payments. The hospital counts by cap status sum to more than the total because hospitals move between the groups over time. The fourth column is the estimation sample: 2011–2024 for hospitals that have been teaching since 1996. Residents are unweighted FTEs. Effective PRA is defined in Equation 3, and its DGME and IME rows are its two components; the GME payment rows are payments actually received. Monetary variables are expressed in 2024 dollars.

3 Stylized Model of the Market for Residents

In this section, I present a stylized model of the market for residents to fix ideas and to illustrate the relationship between caps and how effective GME payments are at encouraging resident training. The model has two parts: a hospital’s choice of residents, which determines demand, and the market for residents, which determines how many residents are trained.

Demand for residents. A teaching hospital i chooses the number of residents R to train, holding fixed its mix of residents across specialties and between residents and fellows. Let $V_i(R)$ be the value of residents to the hospital, net of the non-wage costs of training them, and let $MV_i(R) = V_i'(R)$ be the marginal value of a resident (net of training costs), with $MV_i > 0$ and

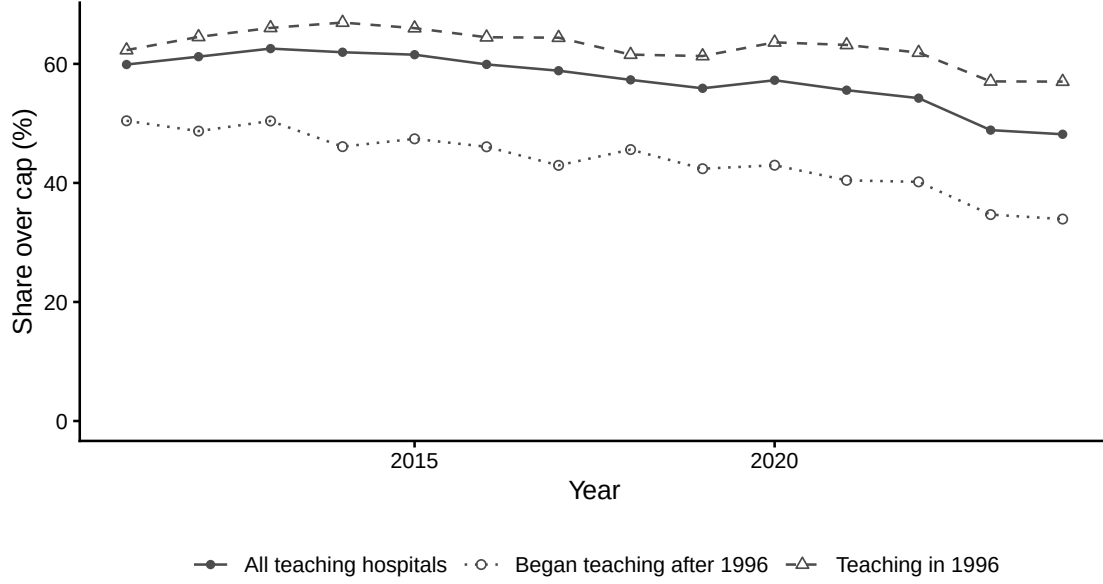


Figure 1: Share of Teaching Hospitals Over Cap, 2011–2024

All teaching hospitals receiving GME payments in each year. A hospital is over-cap if it receives no marginal payments. Hospitals are split by whether they trained residents in 1996.

declining as the hospital adds residents. Residents are paid a wage w , which the hospital takes as given. Medicare pays the hospital p_i for each resident up to its cap K_i , so the hospital's GME payment is $p_i \min(R, K_i)$, where p_i is the effective per-resident amount. The hospital solves

$$\max_R V_i(R) - wR + p_i \min(R, K_i). \quad (4)$$

The marginal payment is p_i below the cap and zero above it, so the payment raises the hospital's marginal return to a resident from $MV_i(R)$ to $MV_i(R) + p_i$, but only up to the cap. The hospital's optimal choice R_i^* therefore satisfies $MV_i(R_i^*) = w - p_i$ if $R_i^* < K_i$ and $MV_i(R_i^*) = w$ if $R_i^* > K_i$. Hospitals with $w - p_i < MV_i(K_i) < w$ choose exactly $R_i^* = K_i$.

Proposition 1 *For a hospital under its cap, resident demand increases with the effective per-resident amount: $\partial \ln R_i^* / \partial \ln p_i > 0$. For a hospital over its cap, resident demand does not depend on the effective per-resident amount: $\partial \ln R_i^* / \partial \ln p_i = 0$.*

The proposition follows directly from the first-order conditions. For an under-cap hospital, a higher p_i lowers the net cost of a resident and moves the hospital down its marginal value curve. For an over-cap hospital, p_i does not enter the first-order condition: it changes the hospital's total payment but not its marginal payment.

Figure 3 illustrates the two cases. The horizontal axis represents the number of residents

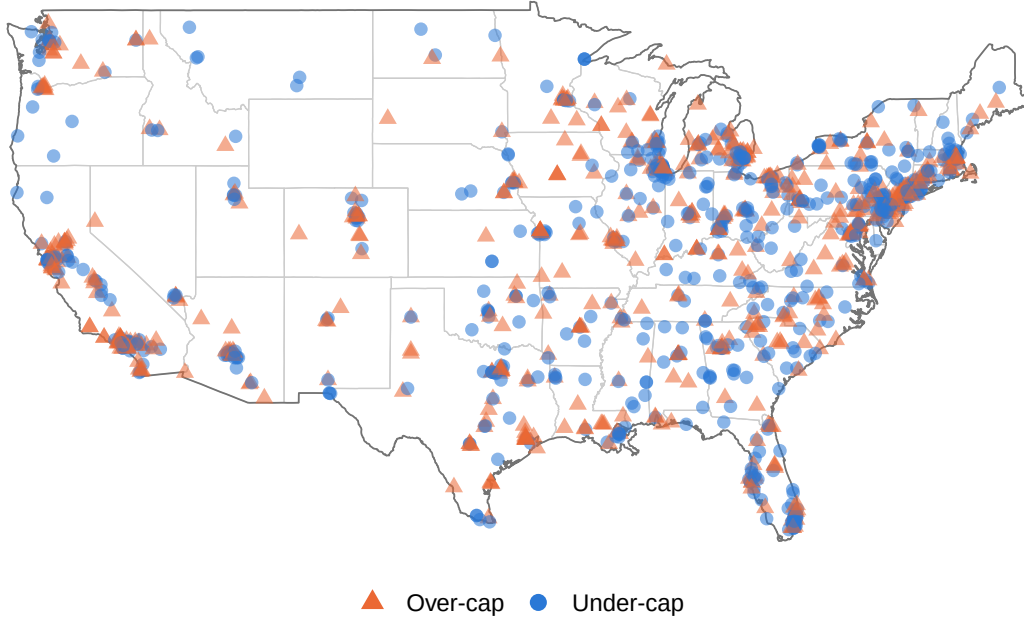


Figure 2: Teaching Hospitals by Cap Status, 2024

All teaching hospitals receiving GME payments in 2024 in the continental United States (1,294 hospitals). A hospital is under-cap if it receives any marginal payments. Each hospital is placed at the center of its ZIP code area, or of its county for the 98 hospitals whose ZIP code has none.

trained. The dashed line is the marginal value of a resident without payments, and the solid line adds the payment, which shifts the curve up by p until the cap and then drops back to the original curve. The hospital trains residents until its marginal value falls to the wage. R_0 is the number of residents that the hospital chooses to train without GME payments, and R^* is the number of residents with GME payments.

On the one hand, if the hospital is over-cap, then its cap has been set below the number of residents that it would have trained regardless of the payments. Its resident count is unchanged by payments, $R^* = R_0$. On the other hand, if the hospital is under-cap, then GME payments do increase its resident count, $R^* > R_0$. Larger per-resident amounts lead to larger increases in its resident count.

Supply of residents. The wage line in Figure 3 is flat because, due to institutional frictions, resident wages do not adjust to clear the market (Gottlieb & Nicholson, 2026; Nicholson, 2003). In the Match, almost all positions are filled while many applicants are unmatched. A hospital can therefore fill as many positions as it chooses at the going wage.

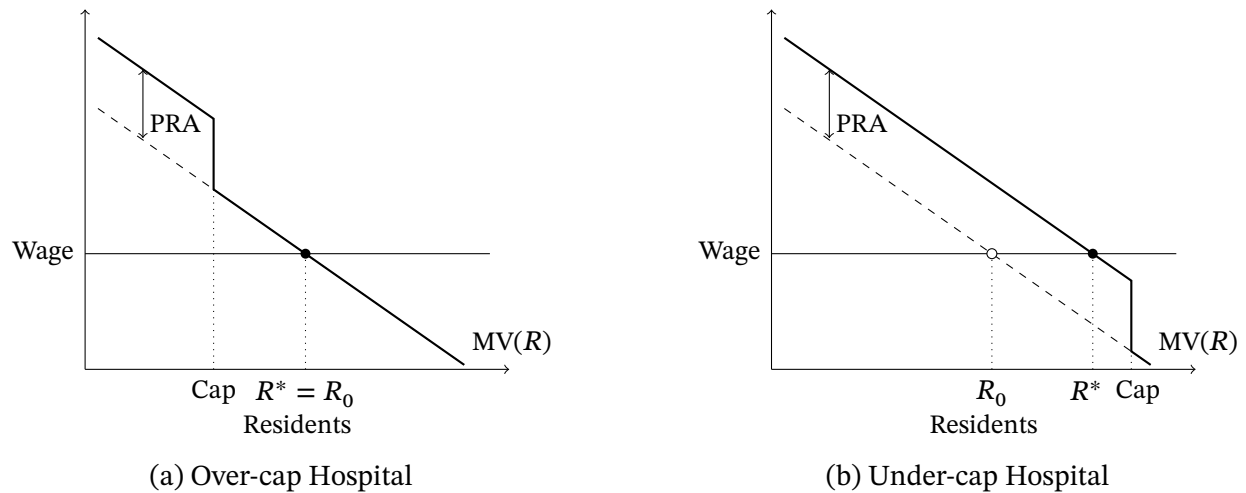


Figure 3: Stylized Model of the Market for Residents

Horizontal line is the fixed wage (w in the text). Dashed line is the marginal value of residents (net of training costs) without GME payments, $MV(R)$. Solid line is with GME payments, which add a per-resident amount (p in the text) up to a cap. The hospital trains residents until the marginal value equals the wage. R_0 is residents without payments, and R^* is residents with payments. Payments have no effect on the over-cap hospital, but they increase residents at the under-cap hospital.

Implications. The model has three implications for the empirical analysis and the policy exercises. First, the elasticity of resident count with respect to the effective per-resident amount should be positive for under-cap hospitals and zero for over-cap hospitals. This is the key prediction that can be tested empirically. Second, as long as there is excess supply, this elasticity measures hospitals' demand response, which translates directly into residents trained. Third, a budget-neutral transfer from over-cap to under-cap hospitals raises the number of residents trained by an amount determined by the under-cap elasticity. Quantifying this elasticity is the first step to designing such a reallocation.

Two caveats apply to this stylized model. First, the model treats the hospital as the decision-maker, whereas residency positions are usually staffed at the program level across different specialties and subspecialties. Thus, the hospital-level choice should be interpreted as an aggregation over program-level choices. Second, the model abstracts away from accreditation requirements and supervisory capacity, which may constrain choices in the short run, so the model should be interpreted as making long-run predictions.

4 Empirical Strategy and Results

4.1 Specification

I estimate the following equation for hospitals i in years t :

$$\ln(\text{Residents}_{it}) = \alpha \text{UnderCap}_{it} + \beta_0 \ln(\text{EffPRA}_{it}) + \beta_1 \text{UnderCap}_{it} \times \ln(\text{EffPRA}_{it}) + \gamma' \mathbf{x}_{it} + \delta_{s(i)t} + \epsilon_{it} \quad (5)$$

where Residents_{it} is resident count, UnderCap_{it} is an indicator for under-cap, and EffPRA_{it} is effective per-resident amount. $\delta_{s(i)t}$ are state-year fixed effects, and $E[\epsilon_{it}] = 0$. The state-year fixed effects absorb anything that varies by state and year, such as health care demand or state policies, so the estimates compare hospitals within the same state in the same year. The exception is the 3 hospitals that are the only teaching hospitals in their states, which are fit exactly by their own fixed effects, so the main regression uses 9,995 of the 10,037 hospital-years.

Observed characteristics $\mathbf{x}_{it}$ consist of logged beds, logged average wage, for-profit and public ownership indicators, medical school affiliation status, and rural location. The wage controls for differences in local labor costs and in the skill mix of the hospital's staff. Ownership, medical school affiliation, and location control for differences in the kinds of hospitals that train residents.

The parameters of interest are β_0 , the elasticity of resident count with respect to GME payments among over-cap hospitals, and $\beta_0 + \beta_1$, the same elasticity among under-cap hospitals. Proposition 1 predicts that $\beta_0 = 0$ and $\beta_0 + \beta_1 > 0$. I center $\ln(\text{EffPRA}_{it})$ at its sample mean, so α is the difference in logged resident count between under-cap and over-cap hospitals at the sample mean of logged effective per-resident amount.

4.2 Identification

Instruments. Under-cap status is mechanically linked to resident count: a hospital that trains fewer residents is more likely to be under-cap. As a result, UnderCap_{it} and $\text{UnderCap}_{it} \times \ln(\text{EffPRA}_{it})$ are correlated with ϵ_{it} . To address this endogeneity, I instrument for them using the change in logged beds since 1996, $\Delta \ln(\text{Beds}_{it})$, and its interaction with logged effective per-resident amount, $\Delta \ln(\text{Beds}_{it}) \times \ln(\text{EffPRA}_{it})$. The present-day level of logged beds is among the controls. The instrument therefore compares two hospitals of the same size today, one of which has grown more since 1996. The hospital that grew more was smaller in 1996, trained fewer residents, and received caps that are smaller relative to its present-day size, making it more likely to be over its caps today.

First stage. Table 2 shows the first stage results. Hospitals with more growth since 1996, and were thus assigned lower caps, are less likely to be under their caps conditional on observed characteristics. At the sample mean of logged effective per-resident amount, a 10% increase in beds since 1996 lowers the probability of being under-cap by 1.9 percentage points, with a coefficient of -0.186 (standard error 0.039).

Table 2: First Stage

	UnderCap (1)	UnderCap $\times$ ln(EffPRA) (2)
$\Delta \ln(\text{Beds})$	-0.186^{***} (0.039)	0.034 (0.028)
$\Delta \ln(\text{beds}) \times \ln(\text{EffPRA})$	0.051 (0.029)	-0.431^{***} (0.061)
F statistic	15.2	24.9
Observations	9,995	9,995

First stage for 2SLS estimation of Equation 5. Each column regresses an endogenous variable on the instruments, logged effective per-resident amount, controls, and state-year fixed effects. $\ln(\text{EffPRA})$ is centered at its sample mean, so the coefficient on $\Delta \ln(\text{Beds})$ alone is its effect at the sample mean of $\ln(\text{EffPRA})$. F statistics test the excluded instruments. SEs clustered by hospital. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

Identifying assumptions. My estimation strategy relies on two key identifying assumptions. I assess each with a placebo test in Section 4.4.

Assumption 1 *Conditional on $\mathbf{x}_{it}$ and fixed effects, effective per-resident amount is correlated with resident count only through the GME payment channel.*

Effective per-resident amounts vary due to policy-assigned per-resident amounts and exposure to Medicare patients. Assumption 1 could fail in two ways. First, policy-assigned per-resident amounts may reflect real and persistent differences in the costs of training residents, so hospitals with higher per-resident amounts may train fewer residents due to higher costs and not GME payments. However, Newhouse and Wilensky (2001) suggest that most variation in policy-assigned per-resident amounts is driven by idiosyncratic accounting practices rather than real costs. Second, and more important given that the IME component is the larger part of the effective per-resident amount, Medicare-intensive hospitals may provide systematically different kinds of care with different staffing needs. The controls in $\mathbf{x}_{it}$ are intended to absorb the most likely differences, and adding more controls for types of care does not change the results (Table B2).

Assumption 2 *Conditional on $\mathbf{x}_{it}$, including present-day hospital size, and fixed effects, change in hospital size since 1996 is correlated with resident count only through the under-cap status channel.*

Intuitively, the instrument compares two hospitals that have the same present-day size, but only one of them grew since 1996. Assumption 2 would fail if hospitals that have grown since 1996 train more or fewer residents for reasons unrelated to their caps. For example, a growing hospital may be expanding into services that require different staffing, or growth may have momentum, so that a hospital that grew since 1996 is still growing and adding programs today. The data do not show momentum: adding growth over the previous five years as a control leaves the estimates unchanged (Table B2). A more specific concern is persistence of residency program size: conditional on present-day size, a hospital that grew since 1996 likely trained fewer residents in 1996 and may have stuck with that size. Such a direct effect of 1996 size on present-day resident counts would load onto α rather than β_1 , the parameter that matters for the policy exercises, as long as the direct effect does not vary with the effective per-resident amount.

4.3 Main Results

Table 3 reports my main results from estimating Equation 5 with two-stage least squares. Column (3), with state-year fixed effects, is my main specification. Among over-cap hospitals, the estimated elasticity of resident count with respect to effective per-resident amount is -0.183 (with a standard error of 0.258). It is not statistically distinguishable from zero ($p = 0.48$). The elasticity among under-cap hospitals is statistically significant and equal to 0.738 (with a standard error of 0.168). The difference between the two elasticities, β_1 , is significant at the 5% level ($p = 0.012$). A 10% increase in effective per-resident amount increases resident counts at under-cap hospitals by 7.4%, while I find no detectable change at over-cap hospitals. These results are consistent with Proposition 1.

Columns (1), with year fixed effects only, and (2), with two-way state and year fixed effects, show that the results do not depend on fixed effect specification. Across the three columns, the over-cap elasticity is between -0.20 and -0.10 and statistically insignificant, and the under-cap elasticity is between 0.71 and 0.76 and significant at the 1% level.

Illustration Figure 4 illustrates the main result in a way that makes the role of the instrument explicit. I residualize growth in logged beds since 1996 on the controls, logged effective per-resident amount, and state-year fixed effects, exactly as in the first stage, so that the residual is

Table 3: Effect of Effective Per-resident Amount on Resident Count

	(1)	(2)	(3)
α : UnderCap	2.604** (0.934)	2.621* (1.033)	2.569* (1.054)
β_0 : $\ln(\text{EffPRA})$	-0.096 (0.249)	-0.201 (0.248)	-0.183 (0.258)
β_1 : UnderCap $\times\ln(\text{EffPRA})$	0.802* (0.361)	0.962** (0.359)	0.921* (0.366)
$\beta_0 + \beta_1$	0.706*** (0.186)	0.761*** (0.168)	0.738*** (0.168)
Year FEs	Yes	Yes	
State FEs		Yes	
State-year FEs			Yes
Observations	10,037	10,037	9,995

2SLS estimates of Equation 5. Dependent variable is logged residents. UnderCap and UnderCap $\times\ln(\text{EffPRA})$ are instrumented with change in logged beds since 1996 and its interaction with $\ln(\text{EffPRA})$. All columns control for logged beds, logged average wage, for-profit and public ownership, medical school affiliation, and rural location. Column (3) drops hospital-years that are alone in their state-year. β_0 is the elasticity for over-cap hospitals, and $\beta_0 + \beta_1$ is the elasticity for under-cap hospitals. $\ln(\text{EffPRA})$ is centered at its sample mean. SEs clustered by hospital. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

the variation in the instrument used for estimation. I then divide hospital-years into quartiles of this residual and, within each quartile k , estimate by ordinary least squares

$$\ln(\text{Residents}_{it}) = \beta^{(k)} \ln(\text{EffPRA}_{it}) + \gamma^{(k)'} \mathbf{x}_{it} + \delta_{s(i)t}^{(k)} + \epsilon_{it}, \quad (6)$$

which is Equation 5 without the under-cap terms: the quartile takes the place of under-cap status. Hospitals in the lowest quartile grew the least relative to their present size and are the most likely to be under their caps. Proposition 1 therefore predicts that the lowest quartiles have the highest $\beta^{(k)}$.

The prediction holds. The elasticity is 0.73 in the lowest quartile and 0.70 in the second, both significant at the 1% level. It falls to 0.24 in the third quartile and to -0.10 in the highest quartile, both statistically insignificant. Resident counts respond to GME payments only when hospitals are likely to receive them at the margin.

4.4 Placebo Tests

Assumptions 1 and 2 cannot be tested directly, since both concern resident counts. The natural benchmark is the quantity of non-resident hospital labor. It is the closest counterpart to the number of residents, and it is determined by the same hospital characteristics, growth, and



Figure 4: Payment Response by Growth Since 1996

Estimation sample. Growth in logged beds since 1996 is residualized on logged current beds, logged average wage, for-profit and public ownership, medical school affiliation, rural location, $\ln(\text{EffPRA})$, and state-year fixed effects, and hospital-years are divided into quartiles of the residual. Points are the OLS coefficients $\beta^{(k)}$ on $\ln(\text{EffPRA})$ from Equation 6, estimated within each quartile with the same controls and state-year fixed effects, plotted against the quartile's mean residualized growth, with 95% confidence intervals from SEs clustered by hospital.

patient mix that could confound the main estimates, while being outside the scope of GME payments. I conduct two tests, one for each assumption.

Placebo test 1. To assess Assumption 1, I estimate Equation 5 with the same instruments and controls, replacing logged resident count with the log of a measure of non-resident hospital labor. GME payments fund residents, not other hospital workers. Thus, under Assumption 1, effective per-resident amount should not affect non-resident labor in either under-cap or over-cap hospitals, so β_0 and β_1 should be zero.

Placebo test 2. To assess Assumption 2, I estimate the reduced form of Equation 5 by ordinary least squares, with the log of non-resident labor as the outcome and with UnderCap_{it} replaced by the instrument, the change in logged beds since 1996:

$$\ln Q_{it} = \alpha^z \Delta_{it} + \beta_0^z \ln(\text{EffPRA}_{it}) + \beta_1^z \Delta_{it} \times \ln(\text{EffPRA}_{it}) + \gamma^{z'} \mathbf{x}_{it} + \delta_{s(i)t}^z + u_{it} \quad (7)$$

where Q_{it} is the placebo measure, Δ_{it} is the change in logged beds between 1996 and year t , and the other terms are as in Equation 5. If growth since 1996 affected hospital labor demand

through channels other than under-cap status, it would show up in α^z or β_1^z . Under Assumption 2, both should be zero.

Results. Table 4 reports both tests for three measures of the quantity of non-resident labor: non-resident workers, non-resident hours, and registered nurses. For Test 1 the table reports the payment coefficients β_0 and β_1 ; for Test 2 it reports the coefficients on the instrument and its interaction, α^z and β_1^z , which are the terms the test concerns. The first row applies the same two tests to resident counts, where both should detect the payment response.

Table 4: Placebo Tests

Outcome	Placebo Test 1		Placebo Test 2	
	β_0	β_1	α	β_1
ln(Residents)	−0.183 (0.258)	0.921* (0.366)	−0.445** (0.137)	−0.267* (0.123)
ln(Non-resident Workers)	0.119 (0.097)	0.018 (0.134)	−0.040 (0.048)	0.003 (0.051)
ln(Non-resident Hours)	0.149 (0.111)	−0.036 (0.147)	−0.063 (0.047)	0.032 (0.053)
ln(Registered Nurses)	0.127 (0.087)	0.206 (0.113)	−0.054 (0.043)	−0.072 (0.047)

Placebo Test 1 estimates Equation 5 with the same instruments, replacing logged resident count with placebos. Placebo Test 2 estimates the reduced form, Equation 7, replacing logged resident count with placebos, and replacing UnderCap_{it} and $\text{UnderCap}_{it} \times \ln(\text{EffPRA}_{it})$ with their corresponding instruments. Non-resident workers and hours are from the cost reports; registered nurses are from the AHA survey. Controls and FEs are as in Column (3) of Table 3. SEs clustered by hospital. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

All three measures pass both tests at the 5% level. In the first test, the smallest p -value across the 6 coefficients is 0.07, and the differential response by cap status, which is the source of the main result, is small for every measure. The one coefficient with a p -value below 0.10 is the interaction for registered nurses, $\beta_1 = 0.206$ (standard error 0.113, $p = 0.07$), which is positive but less than a quarter of the corresponding coefficient for residents. Nurses are the staff most complementary to residents, since a hospital that expands a residency program may require more nursing hours, possibly explaining a small positive interaction. In the second test, neither growth since 1996 nor its interaction with the effective per-resident amount is associated with any of the three measures, and the smallest p -value on these terms is 0.13. These results support both identifying assumptions.

4.5 Robustness Checks

Table B2 in the appendix reports checks on the main specification. In every one, the over-cap elasticity β_0 remains statistically insignificant, and the under-cap elasticity $\beta_0 + \beta_1$ remains positive and statistically significant.

The first two checks add a second measure of hospital size, county population or inpatient days, as a control, with its change since 1996 as an additional instrument. The pattern is unchanged, but the additional instruments weaken the first stage, so these estimates are less precise than the main one.

The next two checks move the threshold that defines under-cap status, since resident counts are measured with error and a hospital just below its cap may in fact be at it. I consider two checks: counting hospitals up to 5% above their caps as under-cap, and counting hospitals at least 5% below their caps as under-cap. The elasticity is larger when the under-cap group is limited to hospitals clearly below their caps, as expected if some hospitals near the threshold are misclassified, but the conclusion does not depend on where the threshold is drawn.

The next check concerns the two caps. A hospital is under-cap if it is below either its DGME or its IME cap, so a hospital with very different caps can be well below one and above the other. Excluding the hospitals whose IME cap is more than 10% above or below their DGME cap leaves the estimates essentially unchanged, although dropping these hospitals weakens the first stage.

The remaining checks vary the controls and the sample. I consider adding controls for intensive care and neonatal intensive care units, trauma center status, system membership, and occupancy. I consider adding bed growth over the previous five years as a control, the momentum check described under Assumption 2. Finally, I consider extending the sample back to 2005 using cost reports in the previous format.

5 Policy Exercises

In this section, I use my estimates to evaluate two changes to the design of GME payments, applied to all teaching hospitals in 2024. The exercises cover all 1,355 teaching hospitals in that year. The 682 under-cap hospitals train 45,407 residents and receive \$9.5 billion in GME payments, and the remaining over-cap hospitals train 91,255 and receive \$13.1 billion.

To conduct these exercises, I interpret my estimates as elasticities of resident demand with respect to effective per-resident amount. This interpretation relies on the presence of excess supply in the market for residents. In the 2024 Match, 44,853 applicants applied for 41,503 positions; almost all positions were filled while 6,678 applicants were unmatched (NRMP, 2024).

Positions, not applicants, are the scarce side of this market, which is what the model assumes. I take the amount of excess supply as given. I state all results as resident FTEs, the quantity the model and the estimates concern. To compare them with the number of unmatched applicants, I convert them into entering positions by dividing by 3.58, the ratio of resident FTEs at all teaching hospitals to the 38,175 applicants who took up a first-year position that year.

I make several assumptions. First, I assume that elasticities of resident demand with respect to effective per-resident amount are homogeneous, so my elasticity estimate, which comes from under-cap hospitals that have been teaching since 1996, applies to all hospitals. Second, I assume that elasticities are constant over the range of changes in effective per-resident amount that I consider, which is a strong assumption because the changes I consider are larger than the variation in the data. I carry the 95% confidence interval of the under-cap elasticity, from 0.41 to 1.07, through the calculations. Third, I assume that over-cap hospitals have zero response to changes in their effective per-resident amounts, as predicted by the stylized model and consistent with my results.

5.1 Removal of Caps

First, as an unrealistic benchmark, I consider removing all caps, allowing currently over-cap hospitals to receive marginal payments equal to their effective per-resident amounts. This is not a small change from their current marginal payment of zero, and a constant elasticity is not defined at zero. I approximate the response linearly around the average effective per-resident amount of under-cap hospitals, \$185,125: each over-cap hospital's resident count rises by the under-cap elasticity multiplied by its effective per-resident amount relative to that average. This approximation is crude, and the resulting figures should be read as an order of magnitude rather than a prediction.

Under this approximation, removing all caps would raise over-cap hospitals' demand for residents by 65,228, or 71% of the residents they currently train, equivalent to about 18,199 first-year positions. Relative to all residents currently trained, this is an increase of 40% in rural areas and 48% in urban areas. If this additional demand were filled, GME payments would rise by about \$17.3 billion per year. Two components make up that cost. Payments for additional residents account for \$13.9 billion. The remaining \$3.4 billion are attributable to the residents these hospitals already train in excess of their caps, who are unsubsidized today and would be fully subsidized without caps.

This demand would be unlikely to be met. The number of additional first-year positions exceeds the number of unmatched residency applicants by a wide margin, so most would go unfilled, and 99% of them would be at urban hospitals, since over-cap hospitals are over-

whelmingly urban. Removing the caps would thus be expensive, would mainly create unfilled positions, and would do little for rural areas.

5.2 Reallocation from Over-cap to Under-cap Hospitals

The second exercise keeps total GME spending fixed. I reduce every over-cap hospital's effective per-resident amount by a common proportion, which lowers their payments without changing their marginal incentives, and use the savings to raise every under-cap hospital's effective per-resident amount by a common proportion. Under-cap hospitals respond by training more residents, which raises their payments further; I solve for the increase in their effective per-resident amounts that exactly exhausts the transfer. Over-cap hospitals' resident counts are unchanged, and each under-cap hospital's expansion is limited by its own cap.

Table 5 reports the results for reductions of 10%, 25%, 50%, and 100%, the last eliminating over-cap hospitals' GME payments altogether. For illustration, I highlight the 25% reduction. This policy transfers \$3.3 billion a year, raises under-cap hospitals' effective per-resident amounts by 24%, and increases the total number of residents by 3,704, a 2.7% increase in residents nationally. The corresponding number of new first-year positions is about 12% of the number of unmatched applicants, so the increase in resident demand can be easily filled, and it would not require any changes in total government spending. The 95% confidence interval of this increase ranges from 2,889 to 4,198.

Table 5: Budget-neutral Reallocation From Over-cap to Under-cap Hospitals

	Δ Over-cap Effective PRAs			
	−10%	−25%	−50%	−100%
Transfer (\$1b)	1.3	3.3	6.5	13.1
Δ Under-cap Effective PRAs	9%	24%	52%	109%
Additional Residents	2,083	3,704	4,842	5,706
Lower End of 95% CI	1,492	2,889	4,066	5,046
Upper End of 95% CI	2,481	4,198	5,212	6,225
Δ Rural Residents	3.1%	5.3%	6.4%	7.7%
Δ Urban Residents	1.5%	2.7%	3.5%	4.1%

All teaching hospitals in 2024. Over-cap hospitals' effective per-resident amounts, and hence their GME payments, are cut by a common proportion, and the savings are used to raise under-cap hospitals' effective per-resident amounts by a common proportion. Budget-neutral after accounting for the increase in under-cap hospitals' resident counts. Each under-cap hospital's expansion is limited by its own cap. Over-cap resident counts are unchanged. Based on the under-cap elasticity and its 95% confidence interval from column (3) of Table 3. Additional residents are resident FTEs. The last two rows give the change in residents as a percentage of all residents at rural and at urban teaching hospitals. Monetary variables are expressed in 2024 dollars.

The other reductions have qualitatively similar effects with varying magnitudes, and the returns diminish sharply as the cut grows. A smaller reduction of 10% increases the number of

residents by 2,083. Doubling the cut from 25% to 50% doubles the transfer but raises the gain only from 3,704 to 4,842, because many under-cap hospitals have already reached their caps. In the limit, eliminating over-cap hospitals' GME payments altogether and redirecting them to under-cap hospitals would raise under-cap amounts by 109% and add 5,706 residents. A 25% reduction therefore achieves more than half of the potential benefits of a 100% reduction, and payment reallocation on its own cannot deliver more than 5,706 additional residents.

Geographic distribution. The additional residents are trained at under-cap hospitals, which are more likely to be rural. Under the 25% cut, resident counts would rise by 5.3% in rural areas and by 2.7% in urban areas, since rural hospitals are more often under their caps. Thus, the increase is proportionally larger in rural areas, which face the most severe physician shortages, although rural hospitals train only 1.7% of all residents, so the gains start from a small base. Whether additional rural training translates into more rural physicians depends on where residents practice after completing their training, which my data do not address. To the extent that physicians tend to practice near where they trained, some of the rural gains would persist beyond residency (McNamara & Pineda-Torres, [2025](#)).

Implementation. A redesign along these lines could be implemented within the existing payment formulas by scaling policy-assigned per-resident amounts and the IME payment multiplier separately for hospitals currently above and below their caps. However, there are some practical challenges that this paper does not address.

The stylized model predicts that over-cap hospitals have zero response to reductions in their GME payments. In practice, large reductions may affect their choices, especially if there are fixed costs, financial frictions, or budgeting practices mechanically tied to total revenue. In the long run, reductions could lead hospitals to stop maintaining programs above their caps. A separate concern is the threshold itself: a permanent policy tied to under-cap status would give hospitals just above their caps an incentive to drop below them, and hospitals just below an incentive to stay there. The results of this paper provide a rough benchmark, and any practical redesign of GME payments should be carefully targeted, phased in, and monitored.

6 Conclusion

Medicare's graduate medical education payments are intended to support the training of residents at a time when the U.S. is projected to face stark physician shortages. However, because of residency caps, many hospitals receive no additional payments for training more residents,

so their payments may not affect their choices. In this paper, I ask whether GME payments would be more cost-effective if they were redirected from over-cap to under-cap hospitals.

My estimates suggest that such a redesign would be effective. Hospitals below their caps respond to GME payments: a 10% higher effective per-resident amount raises their resident counts by 7.4%. For hospitals above their caps, whose marginal payments are zero, I cannot reject a zero response. Growth in hospital size since 1996 separates the two groups in a way that does not depend on how many residents a hospital chooses to train today. I conclude that a 25% reduction in over-cap hospitals' GME payments, redistributed to under-cap hospitals, would add about 3,700 residents with no additional government spending, with proportionally larger gains in rural areas, where physician shortages are projected to be the worst.

An important question for further study is how GME payments shape the mix of specialties that hospitals train. Physician shortages are expected to be especially acute in primary care (HRSA, 2025), but GME payments barely distinguish between specialties: the direct component of GME pays very similar per-resident amounts for primary care and non-primary care residents, and the indirect component, the larger part of the incentive, does not depend on specialty at all. This lack of steering between specialties is only exacerbated by the fact that many hospitals receive zero marginal payments. Understanding how specialty composition responds to payments would help determine whether GME payments can be redesigned to address not only how many residents are trained, but also what kind.

References

- Cahuc, P., Carcillo, S., & Le Barbanchon, T. (2019). The effectiveness of hiring credits. *Review of Economic Studies*, 86(2), 593–626.
- Gottlieb, J., & Nicholson, S. (2026). Physician competition: Entry and substitution. *Journal of Economic Perspectives*, 40(2), 117–142.
- Gottlieb, J., Polyakova, M., Rinz, K., Shiplett, H., & Udalova, V. (2025). The earnings and labor supply of U.S. physicians. *Quarterly Journal of Economics*, 140(2), 1243–1298.
- HRSA. (2025). *Physician workforce: Projections, 2023–2038*. Health Resources and Services Administration.
- Institute of Medicine. (2014). *Graduate medical education that meets the nation's health needs*. National Academies Press.
- Jain, M., Mommaerts, C., & Weaver, J. (2026). *The limits of targeted hiring subsidies: Evidence from the work opportunity tax credit* (Working Paper 35229). National Bureau of Economic Research.

- McNamara, C., & Hussain, T. (2025). Impact of federal funding for graduate medical education on residency program size: Evidence from the Affordable Care Act. *PLOS ONE*, 20(2), e0318626.
- McNamara, C., & Pineda-Torres, M. (2025). Medical residency subsidies and physician shortages. *Journal of Public Economics*, 251, 105494.
- Newhouse, J. P., & Wilensky, G. R. (2001). Paying for graduate medical education: The debate goes on. *Health Affairs*, 20(2), 136–147.
- Nicholson, S. (2003). *Barriers to entering medical specialties*. National Bureau of Economic Research.
- Nicholson, S., & Song, D. (2001). The incentive effects of the Medicare indirect medical education policy. *Journal of Health Economics*, 20(6), 909–933.
- NRMP. (2024). *Results and data: 2024 main residency match*. National Resident Matching Program.
- Saez, E., Schoefer, B., & Seim, D. (2019). Payroll taxes, firm behavior, and rent sharing: Evidence from a young workers’ tax cut in Sweden. *American Economic Review*, 109(5), 1717–1763.

A Data Appendix

Cost report data. The Healthcare Cost Report Information System provides, for each fiscal year, a report file with one record per cost report and numeric and alphanumeric files with one record per worksheet item. Fiscal years refer to the federal fiscal year in which the cost reporting period begins. When a hospital has more than one report in a year, I keep the most recently processed one. Form CMS-2552-96 covers periods beginning before May 1, 2010, and Form CMS-2552-10 covers later periods; for fiscal years 2010 and 2011, which appear on both forms, I use the later form when a hospital-year appears on both. The main analysis uses the later form only; the robustness check that extends the sample to 2005 uses both.

Locating items across forms. The two forms record the same information in different places. Hospital characteristics come from Worksheet S-2, beds, days, and employees from Worksheet S-3 Part I, and wages and hours from Worksheet S-3 Part II. DGME items come from Worksheet E-4 on the later form and Worksheet E-3 Part IV on the earlier form, and IME items from Worksheet E Part A on both. Section 422 slots appear on Worksheet E-3 Part VI of the earlier form and on the main DGME and IME worksheets of the later form. For each item, I map the line and column on one form to the corresponding line and column on the other using the official instructions. On the earlier form, the DGME worksheet changed its layout for periods beginning on or after October 1, 2001, and I rebuild the approved amount under the

rules in force in each period. A small number of reports enter values in the wrong lines, such as per-resident amounts in lines meant for resident counts; I treat such entries as missing.

Reconstructing payments. For DGME, I compute the capped weighted resident count, the three-year rolling average, residents in new programs and residents displaced by closures, the approved amount for primary care and other residents, the approved amount for Section 422 slots, and the Medicare share including Medicare Advantage days net of the nursing and allied health reduction. The rolling average divides by three, except on the earlier form for periods ending before June 30, 2006, when it divides by the number of years with residents. The DGME cap applies in weighted units before October 1, 2022 and in unweighted units from then on; I apply each rule to the periods for which it was in force when hospitals made their decisions. For IME, I back out the DRG payment base from the reported IME payment and resident-to-bed ratio, using the multiplier in effect over the cost reporting period. The multiplier changed several times between 1997 and 2008, as set out in 42 CFR 412.105(d)(3); I use the average over the days of each cost reporting period. For fiscal year 2002 and later, the DRG base backed out in this way matches the DRG payments reported directly on the cost report for more than 95% of teaching hospitals. Before October 2014, the IME payment line includes payments for Medicare Advantage enrollees; I therefore use the sum of the fee-for-service and Medicare Advantage components throughout.

Validation. Reconstructed DGME payments match the payments reported on the cost reports to within 1% for more than 90% of teaching hospitals in most years, and reconstructed IME payments do so for more than 98%. The main exceptions are fiscal years 2019 through 2022, in which many reports were settled retroactively using the fellow-weighting rule adopted in fiscal year 2023.

Inflation adjustment. I convert all monetary variables to 2024 dollars using the average of the monthly consumer price index for all urban consumers over the months of the cost reporting period.

Other data sources. Hospital characteristics from the American Hospital Association Annual Survey are linked to cost reports through the Medicare provider numbers. When several records map to one provider number in a year, I keep the largest facility. The fiscal year 1996 resident counts used to define the sample come from the data released under Section 131 of the Consolidated Appropriations Act of 2021. Rural counties are those with rural-urban

continuum codes of 4 or higher, which correspond to nonmetropolitan counties under the Office of Management and Budget’s delineation.

B Additional Tables and Figures

Table B1: Effective Per-resident Amount vs. Actual Marginal Payments

	Hospital-years	Mean (\$1k)	
		Effective PRA	Actual
Above All Caps	9,419	168.7	0.0
Below Some Caps (Including Sec 422)	2,594	198.2	106.1
Below All Caps (Excluding Sec 422)	4,473	161.2	161.2

Hospital-years in 2011–2024 for all teaching hospitals receiving GME payments; 297 hospital-years with residents but no payments are excluded. A hospital is above all caps if both its DGME and its IME counts are at or above their caps, including Section 422 slots; below some caps if it is above one cap but below the other, or within its Section 422 slots; and below all caps otherwise. A hospital with no cap on record is below all caps. Effective per-resident amounts are defined in Equation 3. Actual per-resident amounts are the marginal payments hospitals receive given observed resident counts and caps: the full effective amount below all caps; the margin from one program only, or the lower Section 422 rates, below some caps; and zero above all caps. Monetary variables are expressed in 2024 dollars.

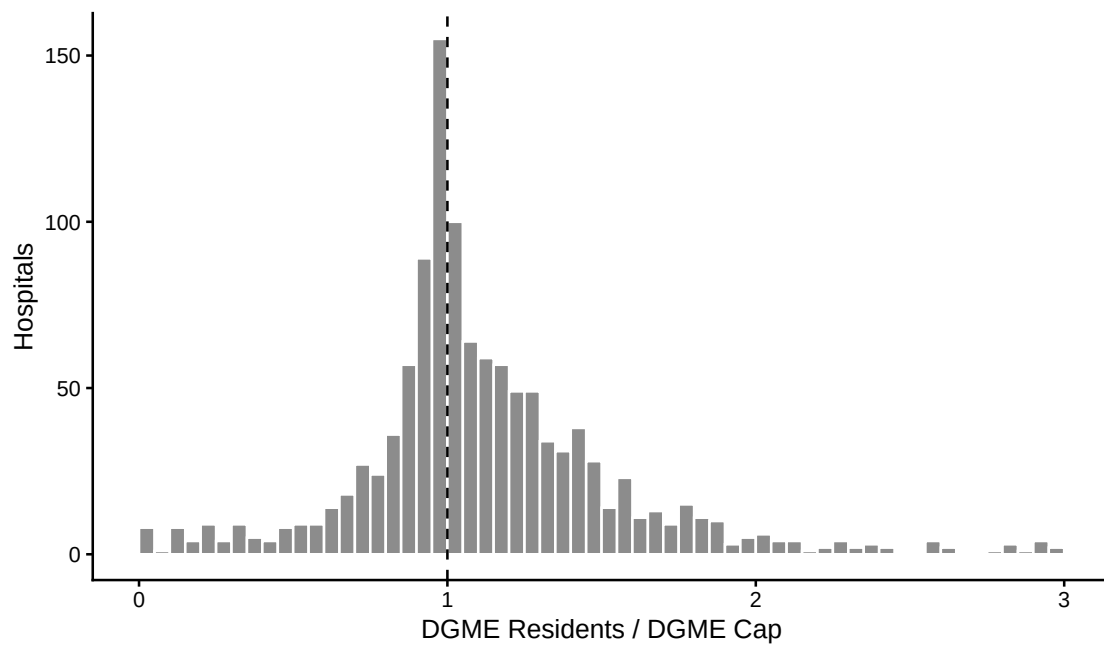


Figure B1: DGME Residents Relative to DGME Cap, 2024

All teaching hospitals with a DGME cap in 2024 (1,223 hospitals). The ratio of residents to the DGME cap, including bonus cap slots from Section 422, in the units in which the cap applies. Dashed line indicates a resident-to-cap ratio of 1. The 57 hospitals (4.7%) with ratios above 3 are not depicted. A hospital above its DGME cap is under-cap in this paper's sense if it is below its IME cap.

Table B2: Robustness Checks

	β_0	$\beta_0 + \beta_1$	F	N
Main Specification	−0.183 (0.258)	0.738*** (0.168)	15.2	9,995
County Pop as Additional Instrument	−0.035 (0.232)	0.709*** (0.151)	8.3	9,920
Inpatient Days as Additional Instrument	0.098 (0.248)	0.475* (0.187)	8.9	9,995
Threshold: 5% Above Cap	−0.347 (0.285)	0.642*** (0.141)	19.2	9,995
Threshold: 5% Below Cap	−0.055 (0.240)	0.951*** (0.261)	11.9	9,995
Exclude Hospitals With Different Caps	−0.170 (0.299)	0.761*** (0.192)	10.2	7,672
Add Hospital Character Controls	−0.204 (0.216)	0.568*** (0.143)	19.6	9,995
Add Bed Growth Over Previous Five Years	−0.172 (0.264)	0.755*** (0.168)	12.1	9,861
Sample From 2005	−0.361 (0.240)	0.643*** (0.189)	13.7	14,570

Each row reports estimates of Equation 5, changing one element of the main specification in column (3) of Table 3. F is the smaller of the two first-stage F statistics. N is the number of hospital-years in the regression. The additional instrument rows add county population or inpatient days as a control, as well as the change since 1996 and its interaction with $\ln(\text{EffPRA})$ as instruments. The threshold rows count hospitals as under-cap if their resident-to-cap ratios are below 1.05 or 0.95 instead of 1. The different-caps row excludes the hospitals whose IME cap differs from their DGME cap by >10%. The character-controls row adds indicators for intensive care and neonatal intensive care units, trauma center status, and system membership, and the occupancy rate. The recent-growth row adds the change in logged beds over the previous five years. The last row starts the sample in 2005, using cost reports in the previous format for 2005–2010. SEs clustered by hospital. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.