Treatment Choices by Altruistic yet Strategic Medical Professionals

Chao Fu* Jesse Gregory† Hsuan-Chih (Luke) Lin‡ Atsuko Tanaka§

April 17, 2026

Abstract

Structuring reimbursements to medical providers is challenging because providers have significantly more information than payers about which treatments are clinically optimal, and most commonly used reimbursement systems create financial incentives for providers to deviate from clinically optimal treatment. We develop and estimate a model of doctors' decisions on hip replacement surgeries using data from Taiwan between 2009 and 2011, a period during which reimbursement schedules underwent several major reforms, including a switch from fee-for-service to Diagnosis-Related Group (DRG) reimbursement and a change in the reimbursement formula under DRG. We analyze how these changes influenced doctors' treatment choices, and use the model estimated from these responses to study the optimal design of reimbursement schedules that align financial incentives with providing the clinically optimal care.

*University of Wisconsin and NBER. cfu@ssc.wisc.edu
†University of Wisconsin and NBER. jmgregory@ssc.wisc.edu
‡National Taiwan University. hclin@ntu.edu.tw
§University of Calgary. atanaka@ucalgary.ca

1 Introduction

The two most common hospital reimbursement systems—fee-for-service and prospective payment—create financial incentives for providers to deviate from clinically optimal treatment choices, and the direction and form of those incentives differ. Under fee-for-service systems, reimbursement for additional treatment often exceeds the marginal cost, creating incentives to overtreat. Under prospective payment systems such as Diagnosis-Related Groups (DRGs), additional treatment spending often does not generate additional reimbursement, creating incentives to undertreat. In addition, because DRG payments depend on the billed diagnosis rather than realized treatment spending alone, they can also encourage selection into more profitable categories.[1] Much of the empirical literature comparing these systems evaluates their relative merits by examining spending and specific health outcomes under each. Such comparisons are informative, but they do not by themselves reveal whether treatment under either system is clinically optimal: that benchmark is not directly observed, because neither fee-for-service nor DRG reimbursement aligns financial incentives with clinically optimal care. Therefore, evaluating and designing reimbursement systems requires recovering not only how providers respond along each distortion margin, but also the distribution of clinically optimal treatment for the underlying patient population.

This paper develops a method to both recover clinically optimal treatment and examine how reimbursement systems move treatment away from it. We apply this method to administrative claims data on all hip replacement surgeries in Taiwan from 2009 to 2011. This setting is particularly useful for our research for several reasons. First, for hip replacement, the physician chooses both a discrete procedure category, which determines the billed DRG code, and a continuous level of non-device treatment spending; our data contain both behavioral margins of interest.

Second, Taiwan's National Health Insurance (NHI) is a mandatory single-payer system that provides universal healthcare coverage for all residents. As such, our data cover the entire distribution of patients, free of demand-side complications arising from patients' sorting into different insurers or insurance statuses; this allows us to better focus on healthcare providers' behaviors.

Third, over this period, Taiwan's NHI introduced substantial changes to reimbursement rules. In 2009, partial hip replacements were reimbursed fee-for-service while total hip replacements received flat DRG payments; in 2010, partial replacements also moved to DRG reimbursement and a lower spending threshold was introduced into both DRG schedules; in 2011, a re-calculation of device spending was introduced, making the lower spending threshold harder to meet. These reforms yield observations of the same clinical decision problem under distinct reimbursement schedules, along with several distinct sources of identifying variation, including changes in relative reimbursement

[1] An extensive empirical literature documents provider responses to prospective payment. Early evidence from Medicare's Prospective Payment System finds large reductions in length of stay and raises the question of whether cost containment comes at the expense of patient welfare (Newhouse and Byrne, 1988; Coulam and Gaumer, 1992; Cutler, 1995). More recent evaluations of mandatory bundled payment programs for joint replacement find spending reductions with limited deterioration in measured quality, though savings often operate through post-acute utilization rather than the index hospitalization (Dummit et al., 2016; Finkelstein et al., 2018; Navathe et al., 2018; Barnett et al., 2019).

across procedure categories, the shift from marginal reimbursement to flat payment, and sharp variation at the DRG thresholds.

Exploiting these institutional features, we first provide design-based evidence that providers respond sharply to these incentive changes, along both the discrete coding margin and the continuous treatment-spending margin. The share of admissions receiving total hip replacement rises from 56% in 2009 to 60% in 2010 and 2011, consistent with movement toward the relatively more profitable billed procedure category.[2] Average spending on partial hip replacements fell after partial replacements moved from fee-for-service to DRG reimbursement, consistent with the elimination of marginal reimbursement for additional treatment intensity. In addition, the distribution of treatment spending under DRG exhibits pronounced bunching at the DRG thresholds, whereas the pre-DRG distribution is smooth at the same spending levels. Next, we interpret these patterns with a structural model in which physicians choose both a procedure type and the treatment intensity by trading off profits against a nonpecuniary cost or "guilt" of deviating from clinically optimal care.

The first contribution of this paper is a framework that uses variation across reimbursement regimes to jointly recover two latent objects: the population distribution of clinically optimal treatment and the distribution of physicians' nonpecuniary costs of over- and under-treatment. We represent the latent need distribution with a dense fixed grid over clinically optimal treatment bundles and estimate the probability mass on that grid by maximum likelihood using an EM routine, so the grid serves as a sieve approximation to an underlying continuous distribution of need. Under each reimbursement regime, the model implies a regime-specific choice kernel or a probability distribution over realized treatments for patients at a given level of underlying need. The observed treatment distribution can therefore be viewed equivalently as a mixture of these regime-specific choice kernels, with mixing weights given by the latent need distribution, and as a policy-specific blurred view of that same distribution. Each payment regime spreads a given level of medical need into a different distribution of realized care, so the empirical problem is to recover a common latent need distribution, together with the provider-side heterogeneity that governs responses to incentives, from several observations of treatment that are distorted by policy-specific financial incentives.

The second contribution of this paper is to embed the variation generated by the reimbursement reforms in a structural framework that separately identifies four distortions of care within a single model: up-coding and down-coding across procedure categories, and over-treatment and under-treatment at the intensive margin. Identification comes from observing the same underlying distribution of patient need under three different payment schedules. The 2009–2010 change in relative reimbursement for partial versus total hip replacement identifies substitution on the procedure-coding margin. The move from fee-for-service to DRG payment identifies the response on the spending margin by changing the return to additional treatment intensity. The post-reform

[2]Strategic reclassification toward more profitable DRG codes—often termed "DRG creep"—has been documented across settings since the early years of prospective payment (Simborg, 1981; Carter et al., 1990; Silverman and Skinner, 2004). International evidence confirms the pattern (Jürges and Köberlein, 2015; Barros and Braun, 2017; Groß et al., 2021). Gowrisankaran et al. (2023) structurally estimate the costs and benefits of coding under Medicare, finding that coding frictions interact with revenue incentives to shape hospitals' coding decisions.

notches then partition that spending response into overtreatment and undertreatment: excess mass just above the reimbursement threshold identifies treatment intensity pushed beyond the clinically appropriate level, while the remaining leftward shift in realized spending away from the threshold region identifies treatment intensity below the clinically appropriate level. This combination of a recovered latent need distribution and cross-regime payment variation allows us to quantify how reimbursement design distorts treatment relative to the clinically optimal benchmark through all three margins.

With these primitives in hand, we design counterfactual reimbursement schedules to curb both wasteful spending and deviation from clinically optimal treatment. More broadly, the paper provides a way to use behavior observed under imperfect reimbursement systems to recover the benchmark needed to evaluate those systems.

Related Literature There is extensive evidence that healthcare providers adjust their behavior in response to financial incentives, altering the intensity and type of care provided (e.g Ellis and McGuire, 1986; Cutler, 1995; Gruber et al., 1999; Einav et al., 2018; Eliason et al., 2018; Parkinson et al., 2019; Gupta, 2021; Alexander and Schnell, 2024) or how care is recorded and reported (Silverman and Skinner, 2004; Dafny, 2005; Sacarny, 2018).

Models of prospective payment characterize distortions on the intensive margin (Ellis and McGuire, 1986; Hodgkin and McGuire, 1994; Clemens and Gottlieb, 2014) but typically take the coding decision as given. Work on upcoding and reimbursement thresholds documents strategic coding responses (Dafny, 2005; Eliason et al., 2018; Geruso and Layton, 2020), but does not jointly model treatment intensity. A related methodological literature uses bunching at budget-set kinks and notches to identify parameters governing responses to financial incentives, originally developed in taxation (Saez, 2010; Chetty et al., 2011; Kleven and Waseem, 2013; Blomquist et al., 2021) and increasingly applied in other areas (see Kleven (2016) for a review). These studies typically require assumptions about the counterfactual distribution that our cross-regime variation relaxes. We combine both margins–procedure coding and intensity of treatment–within a structural model that also recovers the latent distribution of clinically optimal treatment.

Healthcare provider are often modeled as partly "altruistic" in the sense that they do not fully deviate from patients' true needs even if financially incentivized to do so. Such restraint may stem from an internal ethical standard that constrains deviations from appropriate care (McGuire and Pauly, 1991) or from reputational concerns that excessive treatment could diminish future patient demand (Dranove, 1988). Our model draws on a theoretical literature on credence goods and physician altruism. Dulleck and Kerschbamer (2006) characterize conditions under which expert advisors provide honest recommendations when consumers cannot verify whether the recommended treatment was necessary, and Jack (2005) derives optimal payment contracts when the payer cannot observe whether a provider is altruistic. Structural empirical studies support the hypothesis that doctors differ in their degrees of altruism towards patients (Chan et al., 2022; Gaynor et al., 2023a). Building on these insights, our model allows for rich heterogeneity across providers in terms of their

"guilt" when their treatments deviate from patients' needs.

The remainder of the paper proceeds as follows. Section 2 describes the institutional setting. Section 3 documents the data and the empirical patterns around the reform. Section 4 presents the model. Section 5 presents the estimation procedure and discusses identification. Section 6 reports parameter estimates and model fit. Section 7.1 quantifies treatment distortions under the observed reimbursement regimes. Section 9 concludes.

2 Institutional Setting

Taiwan NHI Taiwan operates a universal, single-payer National Health Insurance (NHI) program, introduced in 1995, with enrollment consistently above 99% of residents. The program is financed primarily by payroll-based premiums, with government contributions and modest copayments; low-income households receive full premium subsidies. Because nearly all providers contract with the NHI, reimbursement rules directly shape provider behavior across the system. Medical treatment are reimbursed based on NHI reimbursement points; each point corresponds approximately to 1 New Taiwan dollar (NTD) (1,000 NTD $\approx$ US$33).

Hip Replacement In a total hip replacement, both the femoral head (ball) and the acetabulum (socket) are replaced with prosthetic components. In a partial hip replacement (hemiarthroplasty), only the femoral head is replaced; within this category, procedures differ in surgical approaches and implant types. Accordingly, we classify hip replacement admissions into three procedure categories indexed by $k \in \{1, 2, 3\}$: $k = 1$ (less-invasive partial), $k = 2$ (standard partial), and $k = 3$ (total hip replacement).

2.1 Reimbursement Regimes

During our sample period, reimbursement for hip replacement underwent several major changes; we index reimbursement regimes by $P \in \{1, 2, 3\}$ corresponding to calendar years 2009, 2010, and 2011, respectively. To describe the reimbursement schedules, we introduce a few notations: $Q_P(k)$ is the equipment reimbursement component for procedure type k in regime P; $s \in \mathbb{R}_+$ is non-equipment billed spending, which measures treatment intensity for a given procedure; $\tilde{s} \equiv s + Q_P(k)$ denotes the *total points*.

2.1.1 Regime $P = 1$ (2009: pre–TW–DRG)

In 2009, partial hip replacement cases ($k \in \{1, 2\}$) are reimbursed on a fee-for-service basis, while total hip replacement ($k = 3$) is reimbursed by a fixed amount $s_1^f(3)$, such that

$$W_P(k, s) = \begin{cases} s + Q_P(k) & \text{for } k \in \{1, 2\}, \\ s_1^f(3) & \text{for } k = 3. \end{cases} \tag{1}$$

2.1.2 Regimes $P = 2$ (2010: TW–DRG)

Taiwan's TW-DRG system follows the international DRG template: a fixed, case-based payment per admission within a clinically defined diagnosis category, with the payment amount varying across categories.[3] For hip replacement, the DRG schedule introduced in 2010 includes a lower spending threshold $s_P^{LB}(k)$ below which the hospital receives only fee-for-service reimbursement rather than the full flat payment, and an upper threshold $s_P^{UB}(k)$ above which an outlier add-on payment at rate $r = 80\%$ of costs in excess of the upper threshold is made. This upper-bound provision limits the financial exposure from unusually costly cases while preserving the flat-payment incentive for the bulk of the spending distribution.

Under the TW–DRG, the reimbursement function is

$$W_P(k,s) = \begin{cases} s + Q_P(k) & \text{if } \tilde{s} < s_P^{LB}(k), \\ s_P^f(k) & \text{if } s_P^{LB}(k) \le \tilde{s} < s_P^{UB}(k), \\ s_P^f(k) + r\big(\tilde{s} - s_P^{UB}(k)\big) & \text{if } \tilde{s} \ge s_P^{UB}(k). \end{cases} \tag{2}$$

Within each linear segment, the marginal reimbursement with respect to s is

$$\frac{\partial W_P(k,s)}{\partial s} = \begin{cases} 1, & \tilde{s} < s_P^{LB}(k) \quad \text{(low outlier)}, \\ 0, & s_P^{LB}(k) \le \tilde{s} < s_P^{UB}(k) \quad \text{(inlier / flat DRG)}, \\ r, & \tilde{s} \ge s_P^{UB}(k) \quad \text{(high outlier)}. \end{cases}$$

Within a regime P, $k = 1$ and $k = 2$ share the same TW–DRG threshold parameters.

2.1.3 Regimes $P = 3$ (2011: TW–DRG)

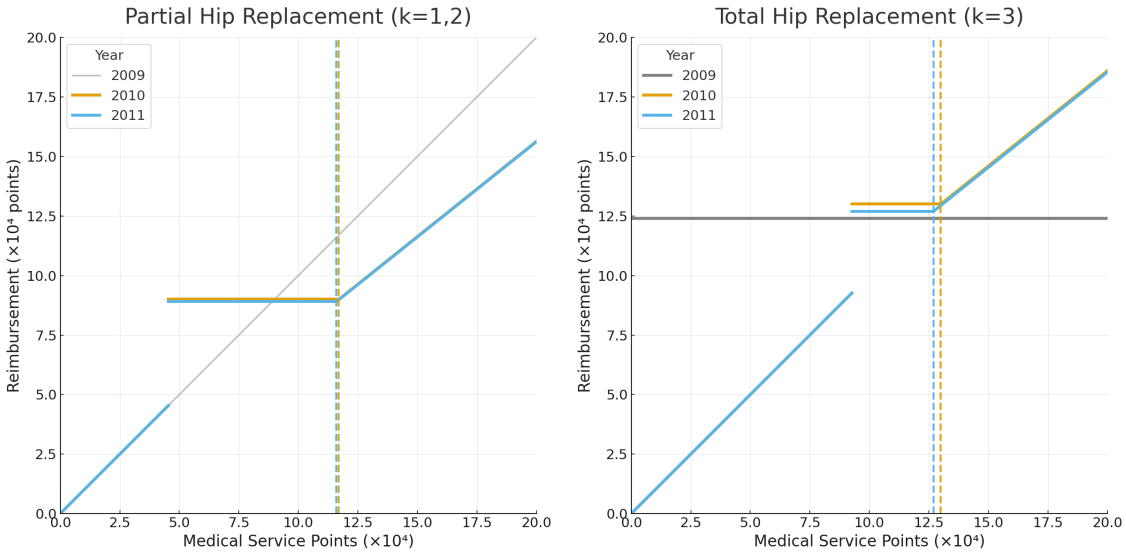
Regime $P = 3$ shares the same reimbursement function (2) under $P = 2$. However, in 2011, NHI re-calculated the cost associated with medical equipment used in hip replacements and updated the reimbursement points $Q_P(k)$, leading to adjustments of all reimbursement parameters accordingly.

Figure 1 illustrates the reimbursement schedules for partial and total hip replacement over all three years. More precisely, Table 1 reports the schedule parameters $(s_P^{LB}(k), s_P^{UB}(k), s_P^f(k), Q_P(k))$ by year and procedure type; all values are in NHI reimbursement points.

3 Data

Our empirical analysis uses de-identified administrative claims from the National Health Insurance Research Database (NHIRD), maintained by Taiwan's National Health Insurance Administration. The NHIRD contains comprehensive billing records for reimbursed medical encounters and links inpatient admissions, outpatient visits, and enrollment information using a unique (anonymized)

[3] For empirical evidence from Taiwan's DRG reform, see, for example, Cheng et al. (2012), Lo et al. (2015), and Chen et al. (2016).



Notes: Reimbursement schedules for partial hip replacement ($k = 1, 2$, left panel) and total hip replacement ($k = 3$, right panel) under Taiwan's National Health Insurance program, 2009–2011. Before 2010, partial hip replacement was reimbursed fee-for-service; total hip replacement received a flat payment. In 2010, both moved to the TW-DRG prospective payment system, which provides a flat case rate between lower and upper spending thresholds and an outlier add-on above the upper threshold. Threshold levels were adjusted in 2011.

Figure 1: TW-DRG Reimbursement Schedules for Hip Replacement

patient identifier. The NHIRD is organized at the individual-month level and includes: (i) inpatient claims (admission-level diagnoses, procedures, and itemized spending), (ii) outpatient and emergency claims (visit-level diagnoses and billed items), and (iii) enrollment files (demographics and premium contributions).

3.1 Sample Construction

From the universe of NHIRD inpatient claims, we select all hospital admissions containing a hip replacement procedure during 2009–2011, identified using ICD-9 procedure codes. We apply two sample restrictions. First, we focus on patients aged 79 and younger. Second, we exclude admissions covered under special funding programs with non-standard reimbursement schedules. Each observation corresponds to an admission. Table 2 summarizes the sample construction. The final analytic sample contains 27,275 hip replacement admissions across 259 hospitals (9,171 in 2009; 8,926 in 2010; 9,178 in 2011).

3.2 Variables

Treatment bundle. For each admission i, we construct the realized treatment bundle $m_i = (k_i, s_i)$, where k_i is a discrete procedure category and s_i is a continuous intensity measure. The three mutually exclusive procedure categories correspond to increasing treatment intensity and equipment cost: $k = 1$ (less-invasive partial hip replacement), $k = 2$ (standard partial hip replacement), and

Table 1: TW–DRG reimbursement schedule parameters (points)

	2009	2010	2011
Lower bound $s_P^{LB}(k)$			
$k=1$ (partial, low equip.)	–	45,325	45,483
$k=2$ (partial, high equip.)	–	45,325	45,483
$k=3$ (total)	–	92,650	92,650
Upper bound $s_P^{UB}(k)$			
$k=1$	–	117,566	115,930
$k=2$	–	117,566	115,930
$k=3$	–	129,736	127,805
Fixed payment $s_P^f(k)$			
$k=1$	–	90,750	88,974
$k=2$	–	90,750	88,974
$k=3$	124,378	129,736	127,805
Equipment component $Q_P(k)$			
$k=1$	10,868	10,868	8,918
$k=2$	47,060	47,060	38,177
$k=3$	52,560	52,560	42,513

Notes: Values are expressed in reimbursement points (often treated as approximately NTD; at the prevailing exchange rate of roughly 30 NTD per US dollar during 2009–2011, one point ≈ US$0.03). Dashes indicate parameters not applicable in 2009 for procedure types not under TW–DRG.

$k=3$ (total hip replacement). Partial versus total replacement is identified by procedure codes; the two partial categories are distinguished by the device/equipment category billed for the admission. The continuous intensity $s_i \in \mathbb{R}_+$ is constructed by taking the difference between the admission-level *total points* $\tilde{s}$ recorded in NHIRD and the category-specific equipment component so that $s = \tilde{s} - Q_P(k)$.

Patient characteristics. From enrollment and claims files, we observe age, sex, and low-income status. We also construct measures of baseline health status from diagnosis codes (e.g., a comorbidity index) and proxy monthly labor income and employment category using premium contributions, because the latter are proportional to pre-tax labor income.

Provider characteristics. We link each admission to hospital characteristics observed in the claims/enrollment data, including hospital tier (medical center, regional hospital, district hospital), ownership (public/private), and teaching status. Our estimation procedure treats the distribution of patients needs in a non-parametric manner and we do not use patient or provider characteristics directly in the estimation; these characteristics are used for descriptive purposes and balance checks.

Table 2: Sample construction

Step	Admissions
All hip replacement admissions, 2009–2011 (NHIRD)	42,835
Exclude age ≥ 80	− 11,519
Exclude special funding programs	− 4,041
Final analytic sample	27,275

Notes: Starting universe is all NHIRD inpatient admissions with an ICD-9 procedure code for hip replacement during 2009–2011. Each row shows the number of admissions dropped at the corresponding restriction step.

3.3 Descriptive Statistics

Table 3 reports summary statistics by year. Panel A describes demographic characteristics of the patient population, including age, sex, and income. Panel B reports patient-weighted hospital characteristics such as hospital type (medical center, regional, or district) and ownership. Panel C reports procedure-type shares and average expenditures. The final column tests for equality of means across the three years. The patient and hospital characteristics are well balanced across years: cross-year differences are not statistically significant at conventional levels for any variable except district hospital share, where the differences, while small in magnitude (21.1%, 21.5%, 20.0%), are statistically significant ($p = 0.037$). This overall stability provides an initial check on the assumption that the composition of underlying medical need is constant over our sample period—an assumption we examine more carefully in Section 3.4.

3.4 Data Patterns around the Reimbursement Reform

The data reveal two salient behavioral responses to the reimbursement reform: a shift in the procedure mix toward total hip replacement, and bunching in the spending distribution at the DRG lower-bound threshold. For each margin we first document pre-reform stability and then show that the pattern emerges sharply after the reform. The pre-reform stability evidence supports our key identifying assumption that the distribution of latent medical need is stable across years (Assumption 1 in Section 5). If medical need were shifting for reasons unrelated to the reform, we would see trends in treatment choices even in the pre-reform period, when reimbursement rules were unchanged. Constant incentives applied to a constant patient pool should produce constant choices. Pre-trends in treatments would not prove that need is shifting (since treatments also reflect incentives), but they would be a red flag.

3.4.1 Procedure Mix

Figure 2 plots the annual fraction of hip replacement admissions that are total hip replacements from 2006–2011. Over 2006–2009 (blue points), the total-hip share fluctuates modestly but shows no systematic trend. In the post-reform years (red points), the fraction of admissions receiving a total hip replacement increases relative to the immediately preceding pre-reform year—from 56% in

Table 3: Summary statistics by year

	2009	2010	2011	All	$p_{09=10=11}$
Panel A. Patient characteristics					
Age (years)	61.4	61.5	61.8	61.6	0.130
Female (%)	52.3	52.9	52.6	52.6	0.733
Avg. monthly income (NTD)	14,544	14,579	15,193	14,772	0.093
Panel B. Hospital characteristics (patient-weighted)					
Medical center (%)	37.9	37.8	38.1	37.9	0.928
Regional hospital (%)	40.2	40.3	41.4	40.6	0.189
District hospital (%)	21.1	21.5	20.0	20.9	0.037
Public hospital (%)	31.1	30.5	30.8	30.8	0.693
Panel C. Procedures and expenditures					
$k = 1$ (partial, low equip.) (%)	1.71	1.85	1.42	–	
$k = 2$ (partial, high equip.) (%)	42.05	38.67	38.93	–	
$k = 3$ (total) (%)	56.24	59.48	59.65	–	
Admissions (N)	9,171	8,926	9,178	27,275	

Notes: Entries are sample means within each year; "All" pools across years. Hospital characteristics are patient-weighted: each admission contributes the characteristics of the hospital where it occurred. Expenditures are expressed in claimed medical points (approximately NTD). Monthly income is computed for the subsample with age < 70. The final column reports the p-value from a one-way F-test of the null that all three year-specific means are equal, assuming independent samples (one-way ANOVA; binary variables use $\hat{p}(1-\hat{p})$ as the within-year variance). For monthly income, the p-value is from an OLS regression with year indicators and standard errors clustered by physician.

2009 to 60% in 2010. This shift is consistent with substitution on the discrete margin: the reform changes relative profitability across procedure categories, and some cases move from partial to total hip replacement, subject to clinical appropriateness and provider disutility from deviating from that benchmark.

3.4.2 Spending Distribution

Conditional on the chosen procedure type, average spending is stable before the reform. Figure 3 reports average reimbursement by year, separately for partial and total hip replacements. In the pre-reform period, total hip replacements are reimbursed by a flat case payment and average reimbursement is essentially constant; partial hip replacements are reimbursed fee-for-service and average reimbursement is similarly stable.

Most importantly for our modeling of the intensive margin, the *full distribution* of spending intensity exhibits no pre-existing discontinuity at the DRG threshold used after the reform. Figure 4 show histograms of non-equipment spending s by procedure type. The dashed vertical line in each panel marks the DRG lower-bound threshold (the eligibility cutoff that becomes financially salient under TW–DRG), expressed in the s-dimension as $s_P^{LB}(k) - Q_P(k)$. In the pre-reform year (2009), the density crosses the threshold smoothly: there is no visible spike and no evidence of a mass

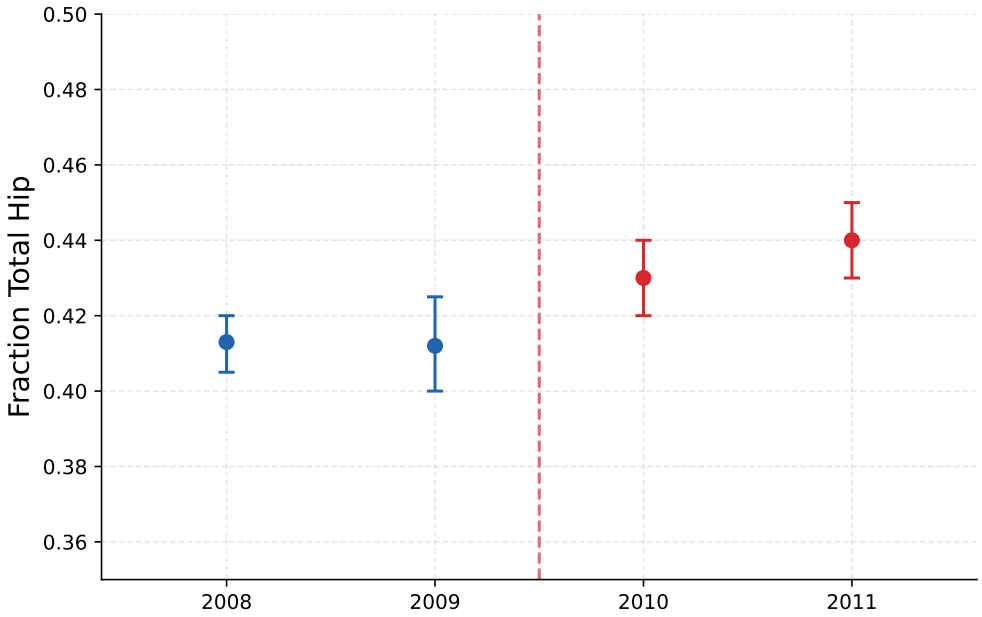


Figure 2: Fraction of hip replacements that are total hip, 2008–2011

Notes: The dashed vertical line marks the introduction of TW–DRG in 2010.

point "pinned" to the lower bound. This absence of pre-existing bunching is informative because it suggests that the threshold is not simply coinciding with a clinical boundary that would generate bunching even without incentives.

Beginning in 2010, the TW–DRG payment system introduces a lower-bound threshold at which reimbursement changes discretely (a *notch*) and a flat-payment region in which marginal reimbursement for additional services is attenuated. Under the notch, providers face a strong incentive to ensure that realized billed intensity clears the lower bound, because crossing the threshold yields a discrete jump in payment. In the post-reform years (2010–2011), the distribution of s displays a sharp discontinuity and a visible spike in the immediate neighborhood above the threshold (Figure 4). The emergence (and strengthening) of this spike is consistent with providers adjusting treatment intensity to reach the DRG eligibility cutoff. Importantly, the spike is not present in the pre-reform year, reinforcing the interpretation that it is induced by the financial notch rather than by a time-invariant clinical rule. Appendix B provides a more detailed decomposition of spending by category, showing that the post-reform reductions are concentrated in categories amenable to within-episode efficiency gains under fixed payment—notably implant choice, drugs, and ward days.

3.4.3 Summary and Implications

Taken together, these patterns establish two facts. First, the absence of pre-trends—in the procedure mix, in mean spending conditional on procedure type, and in the full spending distribution around future DRG thresholds—is consistent with a stable composition of underlying medical need over our sample period, lending credibility to the key identifying assumption of the structural model.

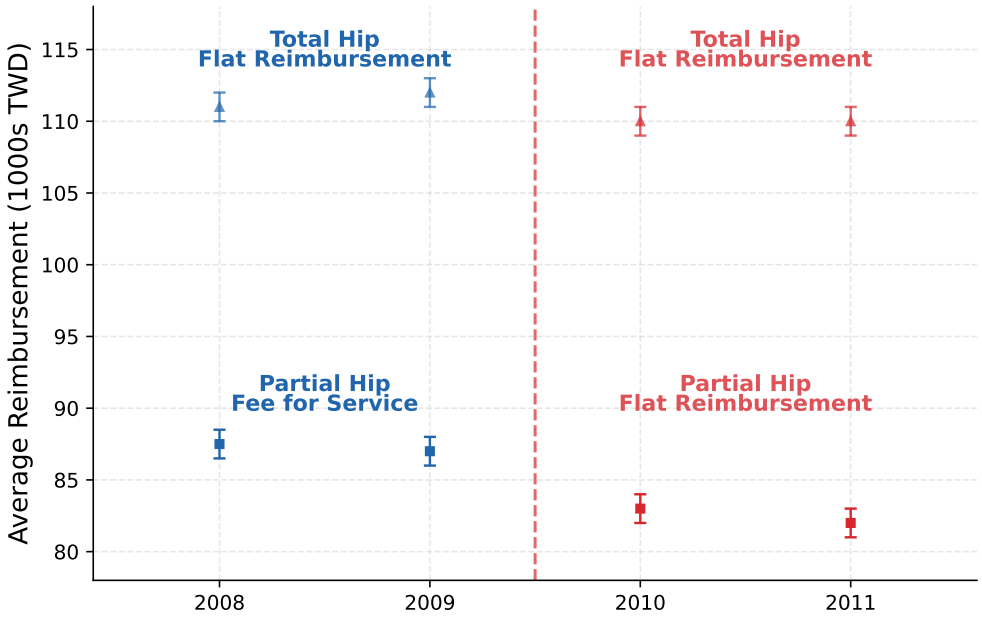


Figure 3: Average reimbursement by procedure type and year

Notes: In the pre-reform years, total hip replacements are reimbursed by a flat payment, while partial hip replacements are reimbursed fee-for-service; under TW–DRG both types are paid by a flat schedule (with outlier provisions).

Second, the sharp post-reform responses—bunching at the DRG notch and a shift in the procedure mix—confirm that providers are responsive to financial incentives on both the intensive margin (spending conditional on procedure) and the extensive margin (procedure choice). Reductions in spending and resource utilization following the TW-DRG reform have been documented across other clinical settings in Taiwan, including hip fractures (Tung et al., 2018), surgical inpatients (Chien et al., 2020), inguinal hernia repair (Yan et al., 2017), and a range of DRG-covered conditions (Hu et al., 2015).

These two facts together motivate the structural model. Since providers respond to incentives, realized treatment choices under *any* regime reflect a mixture of true clinical need and incentive-driven distortions. Neither the fee-for-service era nor the DRG era provides an undistorted picture of underlying medical need: under fee-for-service, the marginal profit from additional spending is positive, creating incentives to overtreat; under the DRG flat payment, the marginal profit from spending is zero (or negative), creating incentives to undertreat. The structural model uses the variation in choices across these regimes, combined with the stability of the underlying patient pool, to disentangle latent medical need from incentive responses.

4 Model

The model includes two ingredients: the population distribution of clinically optimal treatment and preference parameters governing how doctors respond to financial incentives. With these ingredients

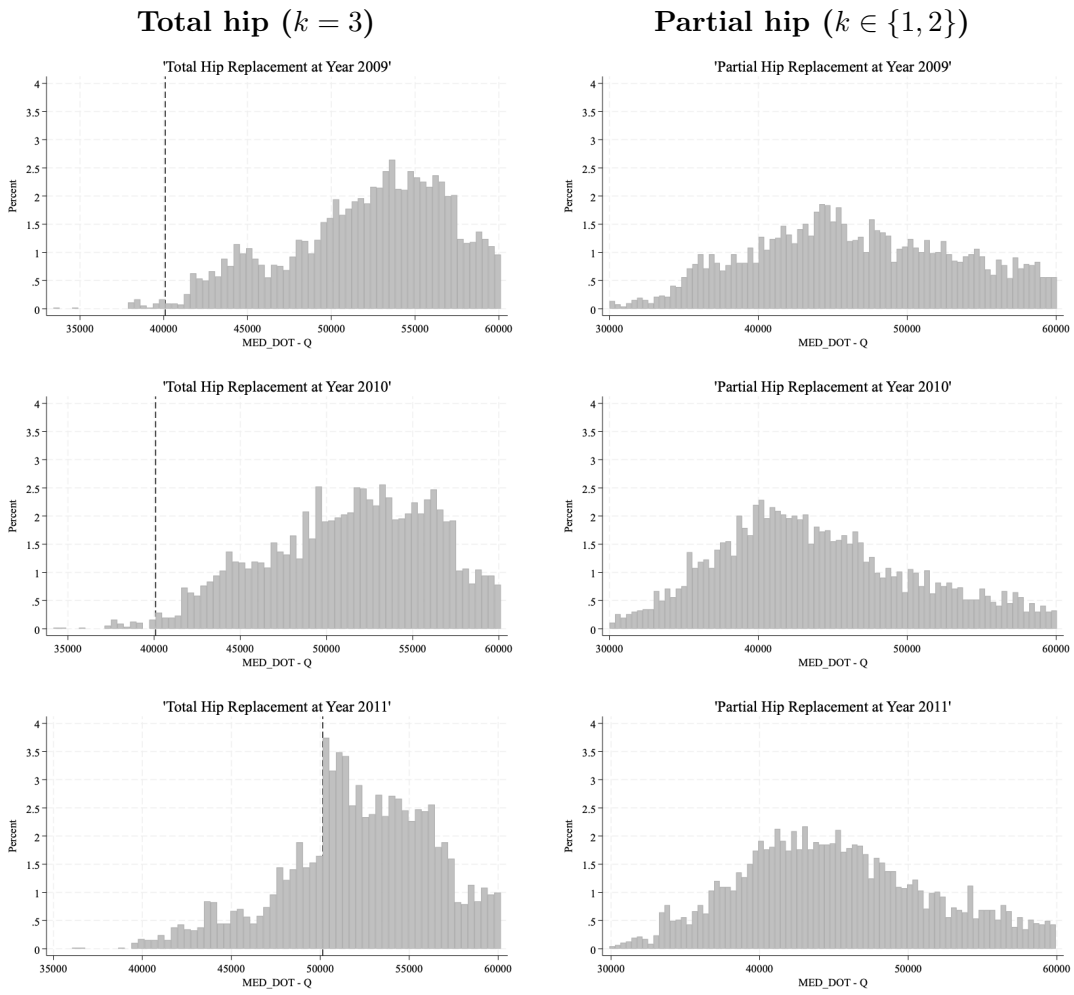


Figure 4: Distribution of non-equipment spending s by procedure type

Notes: The dashed vertical line marks the DRG lower-bound threshold $s_P^{LB}(k) - Q_P(k)$. The 2009 panels use the 2010 threshold for reference. The distribution is smooth at the threshold pre-reform (2009) and exhibits bunching just above the threshold post-reform (2010, 2011).

in hand, we can measure how far realized treatment under any reimbursement regime departs from the clinically optimal benchmark and evaluate alternative payment designs. The basic idea is simple: patients differ in what treatment is clinically optimal for them, and a provider who observes her patient's need trades off profit against the discomfort she feels from deviating from that optimum. The following subsections formalize the tradeoff, define the profit and guilt components, and characterize the resulting treatment choices.

4.1 Primitives

Admissions: There is a distribution $F(m^{\text{opt}})$ of admissions; each needs $m^{\text{opt}} \equiv (k^{\text{opt}}, s^{\text{opt}})$ as its *clinically optimal* treatment, with $k^{\text{opt}} \in \{1, 2, 3\}$ being its optimal discrete procedure category and $s^{\text{opt}} \in \mathbb{R}_+$ being its optimal treatment intensity (non-device spending).

Care Providers: Each admission is matched with a care provider.[4] Observing m^{opt} and her preference shocks (χ, ϵ), a provider chooses treatment $m \equiv (k, s)$.[5] The first preference shock vector $\chi = (\chi^+, \chi^-)$, drawn from a distribution $G(\chi)$ (specified in Appendix A.1, govern a provider's guilt when choosing $s > s^{\text{opt}}$ and $s < s^{\text{opt}}$, respectively. The second preference shock vector $\epsilon = \{\epsilon_1, \epsilon_2, \epsilon_3\}$, associated with the choice of k, is drawn from a Type-1 extreme value distribution with a scale parameter σ_ϵ.

Information: The payer cannot contract on m^{opt} and instead conditions reimbursement on the realized bundle m. Following most of the existing literature, we assume that care providers observe the clinically optimal treatment (m^{opt}). This assumption can be questionable in contexts with high degrees of uncertainties (e.g., predicting heart attacks) and/or large cognitive loads (e.g., emergency care), but are presumably less concerning in our context of hip replacement surgeries.[6] Nevertheless, to the extent that some providers may misjudge and therefore deviate from (m^{opt}), this would be reflected in the model as these providers having low guilt and affect the interpretation of providers' preference estimates.

Key Assumptions: We assume that the distribution of clinically optimal treatments $F(m^{\text{opt}})$ is stable across reimbursement regimes, and that patient-provider pairing is random with respect to unobserved patient need and provider preferences; both are stated formally in Section 5.

[4] We refer to the decision-maker as the "provider". In practice, this can be interpreted as a physician or a physician team.

[5] Given the assumption of independence between m^{opt} and provider preferences, allowing χ, ϵ to be an admission-level draw accommodates both persistent heterogeneity across physicians and case-level factors that shift the effective non-pecuniary cost of deviations.

[6] Mullainathan and Obermeyer (2022) find systematic errors in doctors' judgments in predicting heart attacks, Handel et al. (2026) find that higher cognitive loads lead emergency room physicians to substitute mental deliberation with more numerous but less precise diagnostic actions.

4.2 Provider Preferences

For an admission with true need m^{opt}, a provider with preference shocks (χ, ϵ) will derive the following utility if choosing treatment (k, s) under regime P:

$$u(k, s \mid m^{\text{opt}}, P, \chi) = \underbrace{\pi_P(k, s)}_{\text{profit}} - \underbrace{\Gamma(k, s, m^{\text{opt}}; \chi)}_{\text{guilt}} + \sigma_\varepsilon \varepsilon_k, \tag{3}$$

where $\pi_P(k, s)$ is profit under regime P, $\Gamma(\cdot)$ is a non-pecuniary cost ("guilt") from deviating from m^{opt}, and ε_k is an i.i.d. type-I extreme value shock scaled by $\sigma_\varepsilon > 0$. We define each component of the utility function in turn.

4.2.1 Profits

Under regime P, the payer reimburses the provider $W_P(k, s)$ as described in Section 2.1. Providers incur resource costs. We adopt a constant marginal-cost specification for the continuous intensity component:

$$C(k, s) \;=\; C_0(k) + c\, s, \qquad c \in (0, 1), \tag{4}$$

where c converts billed intensity points into resource costs (in the same units). Under fee-for-service, marginal reimbursement is one point per point of billed intensity (i.e., $\partial W / \partial s = 1$), so the provider's marginal profit from increasing s is $1 - c$. Equivalently, one unit of reimbursement corresponds to c units of resource spending, or a markup of $1/c$ relative to costs. Profits under regime P are therefore

$$\pi_P(k, s) \equiv W_P(k, s) - C_0(k) - c\, s. \tag{5}$$

4.2.2 Guilt

The guilt function Γ penalizes deviations from m^{opt} along both the discrete and continuous margins:

$$\Gamma(k, s, m^{\text{opt}}; \chi) = \delta_{k, k^{\text{opt}}} + \gamma_s(s, s^{\text{opt}}; \chi). \tag{6}$$

Discrete deviation cost. $\delta_{k,k^{\text{opt}}} \geq 0$ is a penalty for choosing procedure type k when the clinically optimal type is k^{opt}. In ordered settings, a convenient restriction and plausible assumption is to rule out non-adjacent jumps by setting $\delta_{k,k^{\text{opt}}} = +\infty$ when $|k - k^{\text{opt}}| > 1$. The discrete-penalty parameters to be estimated is $(\delta_{1,2}, \delta_{2,1}, \delta_{2,3}, \delta_{3,2})$, all restricted to be weakly positive.

Continuous deviation cost (target-centered). We assume a quadratic loss around the patient-specific target:

$$\gamma_s(s, s^{\text{opt}}; \chi) = \chi^+ \mathbf{1}\{s > s^{\text{opt}}\} (s - s^{\text{opt}})^2 + \chi^- \mathbf{1}\{s < s^{\text{opt}}\} (s^{\text{opt}} - s)^2. \tag{7}$$

This parameterization implies that, for a given provider type, the non-pecuniary cost of deviating depends on the *size and sign of the deviation* $s - s^{\text{opt}}$, not on the absolute level of s^{opt}.

4.3 Provider Problem

Let θ collect preference parameters $\{\delta_{k,k^{\text{opt}}}\}$, the shock scale σ_ε, the parameters governing $G(\chi)$, and cost parameters in $C(\cdot)$. Substituting (5) and (6) into (3), the provider chooses (k, s) to solve the following problem:

$$\max_{k\in\{1,2,3\},s\geq 0} \Big\{\pi_P(k,s) - \delta_{k,k^{\text{opt}}} - \gamma_s(s, s^{\text{opt}}; \chi) + \sigma_\varepsilon \varepsilon_k\Big\}. \tag{8}$$

The model delivers two distinct margins of adjustment to a change in reimbursement incentives. First, the provider may substitute across procedure categories k, trading off differences in profitability across categories against the discrete deviation penalties from choosing $k \neq k^{\text{opt}}$. Second, conditional on a procedure category k, the provider selects intensity s by trading off *marginal profits* against the marginal non-pecuniary cost of deviating from the patient-specific target s^{opt}. The provider's problem can be solved via backward induction.

4.3.1 Intensity choice

Given a procedure type k, the provider solves the following problem

$$V_P(k \mid m^{\text{opt}}, \chi; \theta) \equiv \max_{s\geq 0} \Big\{\pi_P(k,s) - \gamma_s(s, s^{\text{opt}}; \chi)\Big\}. \tag{9}$$

Let $s^* \equiv s_P^*(k \mid m^{\text{opt}}, \chi)$ denote the provider's optimal intensity choice for a given k under regime P. Away from kinks/notches/bounds, s^* is interior and satisfies the first-order condition

$$\frac{\partial W_P(k, s^*)}{\partial s} - c \;=\; \frac{\partial \gamma_s(s^*, s^{\text{opt}}; \chi)}{\partial s}. \tag{10}$$

Under the quadratic deviation cost in (7), the right-hand side is piecewise linear:

$$\frac{\partial \gamma_s(s, s^{\text{opt}}; \chi)}{\partial s} = \begin{cases} 2\chi^+(s - s^{\text{opt}}), & s > s^{\text{opt}}, \\ -2\chi^-(s^{\text{opt}} - s), & s < s^{\text{opt}}. \end{cases}$$

Thus (10) can be written as

$$\frac{\partial W_P(k, s^*)}{\partial s} - c = \begin{cases} 2\chi^+(s^* - s^{\text{opt}}), & s^* > s^{\text{opt}}, \\ -2\chi^-(s^{\text{opt}} - s^*), & s^* < s^{\text{opt}}. \end{cases} \tag{11}$$

Let $w_P(k) \equiv \partial W_P(k, s)/\partial s$ denote the marginal reimbursement rate within a given linear segment. When $W_P(k, s)$ is locally linear in s (as in the empirical schedules), (11) implies a simple

mapping from marginal incentives to distortions:

$$s^* = s^{\text{opt}} + \begin{cases} \dfrac{w_P(k) - c}{2\chi^+}, & w_P(k) > c \quad \text{(overtreatment)}, \\ -\dfrac{c - w_P(k)}{2\chi^-}, & w_P(k) < c \quad \text{(undertreatment)}. \end{cases}$$

Figure 5 illustrates the solution: under fee-for-service, doctors shade treatment to the right of s^{opt} (overtreatment); under flat DRG, they shade to the left (undertreatment). Kinks, notches, or bounds in the reimbursement schedule can eliminate an interior solution, in which case the optimum may locate at a schedule threshold (bunching). In our setting, the relevant case is the post-reform notch at the DRG lower bound, which we turn to next.

Bunching at notches. DRG-style payment often introduces *notches*—discrete jumps in reimbursement at policy thresholds defined in terms of total billed points $\tilde{s} = s + Q_P(k)$ (Section 2.1). A notch at $s_P^{LB}(k)$ means that crossing from $\tilde{s} < s_P^{LB}(k)$ to $\tilde{s} \geq s_P^{LB}(k)$ yields a discrete increase in payment.

Notches make boundary solutions natural: for a nontrivial set of (s^{opt}, χ), it is optimal to set s at (or just above) the corresponding threshold in the s-dimension, $\underline{s}_P(k) \equiv s_P^{LB}(k) - Q_P(k)$, because the discrete payment gain can dominate the smooth marginal tradeoff in (11). This generates bunching (excess mass) in the realized distribution of s at the notch point. A lower-bound notch can attract choices from both sides:

- *Overtreatment to reach DRG region.* If s^{opt} lies below $\underline{s}_P(k)$, some providers increase intensity up to $\underline{s}_P(k)$ to secure the higher payment level. This is overtreatment relative to s^{opt}, and it is most pronounced for providers with low χ^+ and for patients with s^{opt} close to the threshold.

- *Bunching from above (undertreatment down to the bound).* If s^{opt} lies above $\underline{s}_P(k)$ in a region where payment is largely fixed, some providers reduce intensity toward $\underline{s}_P(k)$ to save resource costs while preserving the fixed payment. This is undertreatment relative to s^{opt}, and it is most pronounced for providers with low χ^-.

In practice, computing $s_P^*(k \mid m^{\text{opt}}, \chi)$ in the presence of a notch requires evaluating both the interior tangency solution from (11) and the notch-point option $\underline{s}_P(k)$, then selecting whichever yields higher utility. Figure 5 illustrates bunching at the lower bound alongside the overtreatment and undertreatment mechanisms from Section 4.5.1.

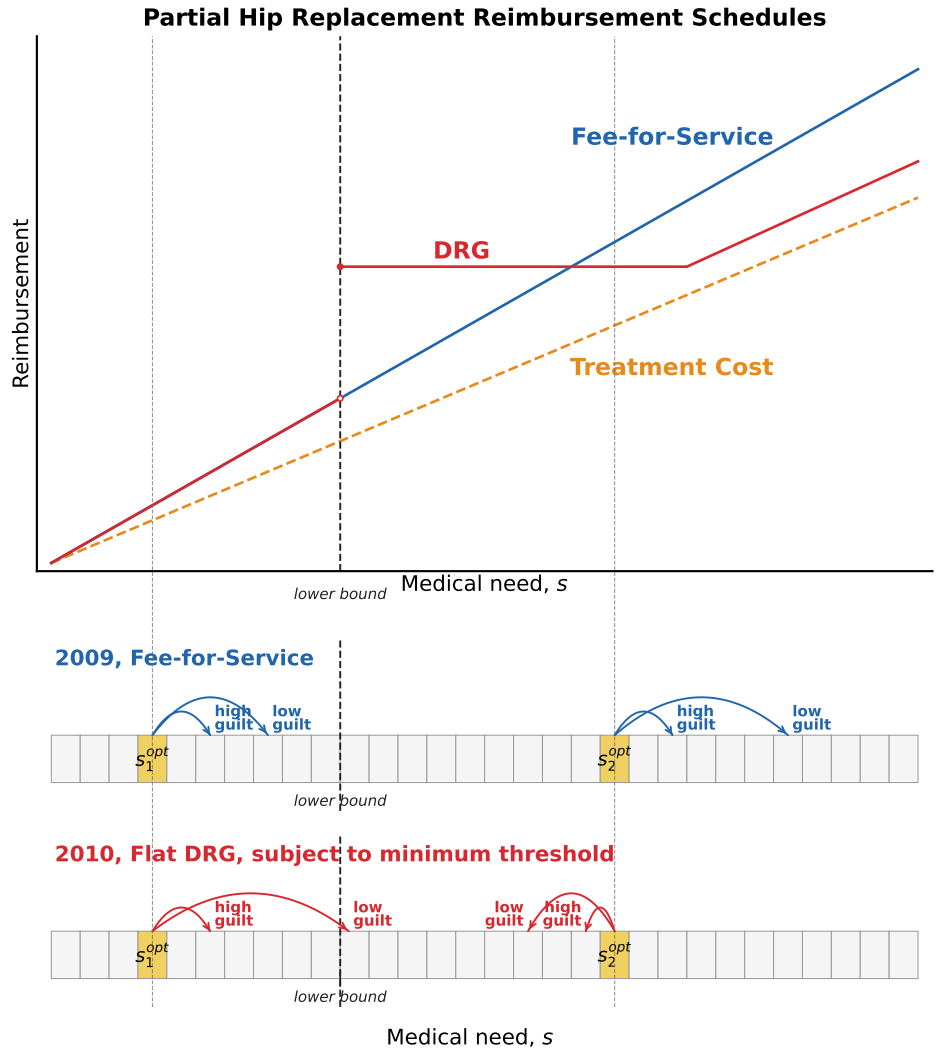


Figure 5: Reimbursement schedules and the blurring of medical need

Notes: Top panel: Fee-for-service (blue) reimburses at slope 1; the DRG schedule (red) matches fee-for-service below the lower bound, then jumps to a flat payment before returning to fee-for-service at a reduced slope. *Bottom panels*: For two illustrative patients with optimal treatments s_1^{opt} and s_2^{opt}, curved arrows show how providers with high versus low guilt shift the chosen treatment away from s^{opt}—rightward (overtreatment) under fee-for-service, leftward (undertreatment) or toward the lower bound under DRG. The observed distribution of treatments under each regime is therefore a "blurred" version of the latent distribution of s^{opt}, with the blurring kernel determined by the s^{opt}-specific mapping from guilt types to treatment choices under the prevailing reimbursement schedule. Treatment cost (orange, dashed) reflects within-treatment costs only, excluding the cost of the device.

4.3.2 Discrete choice of procedure type

The choice of k solves the following problem:

$$\max_{k\in\{1,2,3\}} \left\{ V_P(k \mid m^{\text{opt}}, \chi; \theta) - \delta_{k,k^{\text{opt}}} + \sigma_\varepsilon \varepsilon_k \right\}, \tag{12}$$

where $V_P(k|\cdot)$ is given by Equation (9). Changes in the reimbursement schedule can therefore induce substitution in k whenever the resulting change in relative profitability is large enough to overcome the discrete deviation penalty from selecting $k \neq k^{\text{opt}}$ (and the idiosyncratic taste shocks). Because intensity is re-optimized within each candidate k, switching k changes both the level and the slope of profits in s (and the location of any relevant thresholds), and therefore changes the choice of s.

4.4 Aggregating Individual Choices

Under fee-for-service, $w_P(k) = 1$, so marginal profit from increasing intensity is $1 - c > 0$. The model predicts $s^* > s^{\text{opt}}$ for most providers—overtreatment—with larger distortions for providers with lower χ^+. Under flat DRG, $w_P(k) = 0$, so marginal profit is $-c < 0$, pushing intensity below s^{opt} away from threshold regions, especially for providers with lower χ^-. DRG notches create additional localized incentives to bunch at policy cutoffs, even when the latent distribution of s^{opt} is smooth.

Because providers differ in χ and procedure choice includes idiosyncratic components, patients with the same clinically optimal bundle m^{opt} may receive different realized treatments m. Each reimbursement regime therefore maps the latent distribution of medical need into a different distribution of observed treatments. Formally, let $K_P(m \mid m^{\text{opt}}; \theta)$ denote the model-implied conditional distribution of realized treatment $m = (k, s)$ given latent medical need $m^{\text{opt}} = (k^{\text{opt}}, s^{\text{opt}})$ under regime P, after integrating over (χ, ϵ):

$$K_P(m \mid m^{\text{opt}}; \theta, P) = \iint I(m \mid m^{\text{opt}}, \chi, \epsilon; \theta)\, \mathrm{d}G(\chi)\, \mathrm{d}L(\epsilon). \tag{13}$$

We refer to K_P as the *policy-specific choice kernel*.

Let $f_P(m)$ denote the induced distribution of realized treatments under regime P, and let $f(m^{\text{opt}})$ denote the (latent) distribution of clinically optimal treatments in the population. The model implies the mixture relationship

$$f_P(m) = \int K_P(m \mid m^{\text{opt}}; \theta)\, f(m^{\text{opt}})\, \mathrm{d}m^{\text{opt}}, \qquad P \in \{1, 2, 3\}. \tag{14}$$

Equation (14) is a "forward" mapping from latent need to observed treatments: the distribution of clinically optimal treatments, together with provider behavior under regime P, implies the observed treatment distribution in that regime (see Figure 5).

5 Estimation and Identification

Known Parameters. The reimbursement schedule $W_P(\cdot)$ is treated as observed (policy-given), so schedule features such as thresholds, fixed payments, and outlier add-on rates are known. When discretizing/simulating the likelihood near kink/notch points, we may spread point-mass choices over a small interval of width h for numerical stability; h is a tuning parameter set ex ante (not

estimated). We also treat the marginal resource-cost parameter c in (4) as fixed (calibrated), so that fee-for-service corresponds to "one reimbursed point per c units of spending" (markup $1/c$).

The distribution $F(m^{\mathbf{opt}})$. We approximate the latent distribution $f(m^{\text{opt}})$ with J support points $\{m_j^{\text{opt}}\}_{j=1}^J$ and associated weights $\{w_j\}_{j=1}^J$:

$$f(m^{\text{opt}}) \approx \sum_{j=1}^{J} w_j \, \delta_{m_j^{\text{opt}}}(m^{\text{opt}}), \qquad w_j \geq 0, \quad \sum_{j=1}^{J} w_j = 1. \tag{15}$$

The support points can be drawn once from a proposal distribution $g(m^{\text{opt}})$ or chosen as a fixed grid. The key objects are the weight vector $w = (w_1, \ldots, w_J)$: different weights correspond to different candidate latent distributions, holding the support points fixed.

Parameters to be estimated. Letting θ be the vector of parameters governing the distribution of provider preferences: $(\delta, \sigma_\epsilon)$ and the parameters governing $G(\chi)$, the full set of parameters to be estimated is $\{\theta, w\}$.

5.1 Estimation

The estimation problem has a distinctive structure: the preference parameters θ and the distribution of clinically optimal treatments $f(m^{\text{opt}})$ must be recovered jointly. The preference parameters govern how physicians distort treatment in response to financial incentives; the need distribution is the policy-invariant object against which reimbursement systems are evaluated. Neither is observed directly, and with data from a single reimbursement regime they would not be separately identified, since any change to provider behavior could be rationalized by a change in the underlying patient population. Our approach exploits the fact that three reimbursement regimes generate three differently distorted views of the same latent distribution: the regimes change the mapping from need to treatment but not the need distribution itself. This cross-regime variation allows us to disentangle the two objects using what we call a *cross-regime deconvolution estimator*: an alternating procedure that iterates between (1) a minimum-distance step, which updates θ by matching moments of the observed treatment distributions, and (2) an EM reweighting step, which recovers $f(m^{\text{opt}})$ nonparametrically by deconvolving the observed treatments through the model-implied choice kernels.

Recall that $K_P(m \mid m^{\text{opt}}; \theta)$ denotes the probability that a provider chooses treatment m given medical need m^{opt} under regime P, and that the observed treatment distribution satisfies the mixture equation (14). The cross-regime strategy relies on two assumptions.

Assumption 1. *Stable distribution of clinically optimal treatments:*

$$f(m^{opt} \mid P = 1) = f(m^{opt} \mid P = 2) = f(m^{opt} \mid P = 3) \equiv f(m^{opt}).$$

Under Assumption 1, the same latent distribution generates the observed treatment choices in all three regimes, with regime variation entering only through K_P. Cross-regime comparisons therefore identify causal responses to incentive changes—including the bunching elasticity at notch points and the average shift in spending from fee-for-service to DRG—without further restrictions. To recover the full distribution of clinically optimal treatment, however, we need the preference parameters estimated from bunching and other local variation to be portable across the distribution of patient need. This is the role of the second assumption:

Assumption 2. *Random assignment of provider type (no sorting on latent need): For each admission, the provider preference shocks (χ, ϵ) are independent of the patient's latent clinically optimal treatment m^{opt}: $(\chi, \epsilon) \perp\!\!\!\perp m^{opt}$.*

Independence ensures that the choice kernel K_P estimated from one part of the need distribution—for instance, from bunching behavior near a threshold—applies throughout the support of $f(m^{\text{opt}})$, so the deconvolution in (14) is well-posed. The EM reweighting step uses an iterative algorithm algebraically identical to the expectation-maximization update for a finite mixture model (Train, 2008). The following subsections describe each step, then collect them into the full estimation algorithm.

5.1.1 Preliminaries

Both steps of the estimator operate on a common discrete approximation to the latent distribution and a common procedure for computing the model-implied choice probabilities. We describe these shared building blocks before turning to the two steps themselves.

Support-point representation. We approximate the latent distribution $f(m^{\text{opt}})$ with J support points and associated weights, specified in Equation (15).

Computing the choice distribution $K_P(m \mid m^{\mathbf{opt}}; \theta)$. Both steps of the estimator require evaluating $K_P(m \mid m^{\text{opt}}; \theta)$: the minimum-distance step uses it to compute model-implied treatment distributions, and the reweighting step uses it to form predicted likelihoods. For a given θ and candidate m^{opt}, we solve the provider problem (8) and integrate their choices to compute $K_P(m \mid m^{\text{opt}}; \theta)$ as in (13).[7]

5.1.2 Minimum-Distance Step: Updating θ

Given an estimate $\widehat{f}(m^{\text{opt}}; \theta)$, the minimum-distance step chooses θ to bring the model's predicted treatment distributions into alignment with the data. We form the model-implied distribution for each regime by mixing K_P over the support points:

$$\widehat{f}_P(m; \theta) \equiv \sum_{j=1}^{J} K_P(m \mid m_j^{\text{opt}}; \theta)\, w_j, \qquad P \in \{1, 2, 3\}. \tag{16}$$

[7] In practice, boundary solutions imply point masses in the model-implied distribution of s. When matching binned histograms, we optionally spread such mass over a narrow interval above a threshold for numerical stability.

We target two sets of moments for each regime P: (i) the marginal probability of each procedure type k, $\Pr(k)$, and (ii) the conditional distribution of intensity s given k, summarized by bin shares, $\Pr(s \in \mathcal{B}_b \mid k)$. The model-implied analogs, using the support-point approximation, are

$$p_P(k;\theta,w) = \sum_{j=1}^{J} w_j \cdot \Pr\big(k_i = k \mid m_j^{\text{opt}}, P; \theta\big),$$

$$h_{P,k}(b;\theta,w) = \frac{\displaystyle\sum_{j=1}^{J} w_j \cdot \Pr\big(k_i = k,\, s_i \in \mathcal{B}_b \mid m_j^{\text{opt}}, P; \theta\big)}{p_P(k;\theta,w)},$$

where each probability is computed from $K_P(\cdot \mid m_j^{\text{opt}};\theta)$. Stacking $\{p_P(k;\theta,w)\}$ and $\{h_{P,k}(b;\theta,w)\}$ into $M_{\text{model}}(\theta)$, and their empirical counterparts (detailed in Appendix C) into $\widehat{M}_{\text{data}}$, we choose θ to minimize

$$\hat{\theta} = \arg\min_{\theta} \ \big(\widehat{M}_{\text{data}} - M_{\text{model}}(\theta)\big)^{\top} W \big(\widehat{M}_{\text{data}} - M_{\text{model}}(\theta)\big), \tag{17}$$

for a positive semidefinite weighting matrix W.

5.1.3 Reweighting Step: Recovering $f(m^{\text{opt}})$ for a given θ

The reweighting step recovers $f(m^{\text{opt}})$ for a fixed θ. Given the support-point representation (15), we derive the model-implied likelihood and posterior responsibilities, then describe the iterative weight update.

Microdata likelihood and posteriors Given weights $w^{(q)} = \{w_j^{(q)}\}_{j=1}^{J}$ at iteration q, the model-implied likelihood for observation i is

$$\widehat{f}^{(q)}(m_i \mid P_i;\theta) \equiv \sum_{j=1}^{J} K_{P_i}(m_i \mid m_j^{\text{opt}};\theta)\, w_j^{(q)}. \tag{18}$$

Given the predicted likelihood, we compute the posterior responsibility — the probability that observation i's latent m^{opt} equals support point j under the current weights — by Bayes' rule:

$$\tau_{ij}^{(q)} \equiv \Pr(m^{\text{opt}} = m_j^{\text{opt}} \mid m_i, P_i; \theta, w^{(q)}) = \frac{K_{P_i}(m_i \mid m_j^{\text{opt}};\theta)\, w_j^{(q)}}{\sum_{j'=1}^{J} K_{P_i}(m_i \mid m_{j'}^{\text{opt}};\theta)\, w_{j'}^{(q)}}. \tag{19}$$

Weight update The posterior responsibilities $\tau_{ij}^{(q)}$ determine how the weights are updated. Starting from an initial $w^{(0)}$, we iterate the update

$$w_j^{(q+1)} = \frac{1}{N} \sum_{i=1}^{N} \tau_{ij}^{(q)}. \tag{20}$$

Equation (20) has a direct interpretation: $\tau_{ij}^{(q)}$ is the posterior probability that admission i's medical need came from support point j, given the observed treatment m_i and current weights. The new weight $w_j^{(q+1)}$ is simply the average of these posteriors across admissions, which estimates the unconditional probability that a randomly drawn patient has need m_j^{opt}. At convergence, this self-consistency condition holds exactly: the prior probability assigned to support point j equals the average posterior probability after observing treatments, $w_j = \frac{1}{N}\sum_{i=1}^{N} \tau_{ij}(w)$.

5.1.4 Full Estimation Algorithm: Cross-Regime Deconvolution

The two steps combine into the following algorithm.

1. **Initialize.** Construct support points $\{m_j^{\text{opt}}\}_{j=1}^{J}$ (drawn from a proposal distribution or fixed on a grid). Set $w_j = 1/J$ for all j and choose an initial $\theta^{(0)}$.

2. **Minimum-distance step.** Given the current weights $\{w_j\}$, update θ by minimizing the distance between the model-implied and empirical moments via (17).

3. **Reweighting step.** Given the updated θ, recover the distribution of clinically optimal treatment most jointly consistent with the three observed treatment distributions. Iterate until convergence:

   a. Compute predicted likelihoods $\widehat{f}^{(q)}(m_i \mid P_i; \theta)$ using (18).

   b. Compute posteriors $\tau_{ij}^{(q)}$ using (19).

   c. Update weights $w_j^{(q+1)}$ using (20).

   d. If $\max_j |w_j^{(q+1)} - w_j^{(q)}| < \text{tol}$, proceed to step 4. Otherwise increment q and return to step 3a.

4. **Check convergence.** If θ has stabilized, stop. Otherwise return to step 2.

5.2 Identification

Our identifying variation comes from comparing the distribution of realized treatment choices (k, s) across the three reimbursement regimes. The central assumption is that the distribution of clinically optimal treatments $f(m^{\text{opt}})$ is stable across regimes—changes in the observed distribution of choices reflect changes in provider incentives rather than changes in underlying patient need (Assumption 1). Random assignment of patients to providers ensures that provider heterogeneity in guilt parameters does not vary systematically with patient need.

The upcoding margin is identified from changes in procedure-type shares across regimes. Moving partial hip replacements from fee-for-service to flat DRG in 2010 reduced the profitability of partial relative to total replacements, creating an incentive to upcode from partial to total. The observed rise in the total replacement share—from 56% in 2009 to 60% in 2010—measures the strength of this response, after accounting for the stable distribution of clinically optimal procedure types.

The continuous intensity margin is identified from distributional shifts in spending s across regimes. The switch from fee-for-service to flat DRG changes the direction of the financial incentive for incremental spending. The resulting shift in the distribution of s identifies the combined strength of over- and undertreatment responses, but does not by itself separate the two, since both contribute to the observed change.

Our approach to the notch regions builds on the bunching-based identification literature. Kleven and Waseem (2013) show how excess mass at notch points identifies structural elasticities, and Blomquist et al. (2021) demonstrate that with a single budget set the bunching probability does not separately identify the behavioral parameter and the shape of the counterfactual distribution. Bertanha et al. (2023) develop further nonparametric refinements. In our setting, this underidentification problem does not arise because we observe the same population under three reimbursement regimes, one of which (fee-for-service in 2009) has no notch. The pre-reform regime still distorts treatment through the overtreatment incentive inherent in fee-for-service, so it does not directly reveal the undistorted counterfactual distribution. However, it does provide a spending distribution that is free of bunching, and because we simultaneously estimate the behavioral parameters governing the response to each regime's incentives, the combination is sufficient to separately identify the counterfactual density and the behavioral responses.

The notches introduced into the DRG schedule in 2010 and 2011 provide the key to separating them. A notch creates a localized incentive to raise spending above a threshold to qualify for a higher flat payment. Providers with low overtreatment guilt find it optimal to cross the threshold even when doing so requires spending above the patient's clinically optimal level, generating excess mass just above the threshold. The extent of this bunching identifies overtreatment responsiveness χ^+ directly. Undertreatment responsiveness χ^- is then identified as the residual: away from the notch regions, the extent to which the left tail of the spending distribution expands under DRG relative to fee-for-service identifies the distribution of χ^-, after the overtreatment component has been pinned down by bunching.

Identification of the latent distribution $f(m^{\text{opt}})$ follows from the mixture structure of the model. For each regime P, the model implies

$$f_P(k,s) = \int K_P\big((k,s) \mid m^{\text{opt}};\theta\big)\, f(m^{\text{opt}})\, dm^{\text{opt}},$$

where the kernel K_P encodes how providers respond to the reimbursement schedule. Observing three regime-specific choice distributions generated by three different kernels, while $f(m^{\text{opt}})$ is stable by Assumption 1, jointly identifies both θ and $f(m^{\text{opt}})$. Having identified θ by matching moments across all three regimes, $f(m^{\text{opt}})$ is identified as the solution to the deconvolution of the observed choice distributions through the regime-specific kernels, which we implement with the EM reweighting step described in Section 5.1.

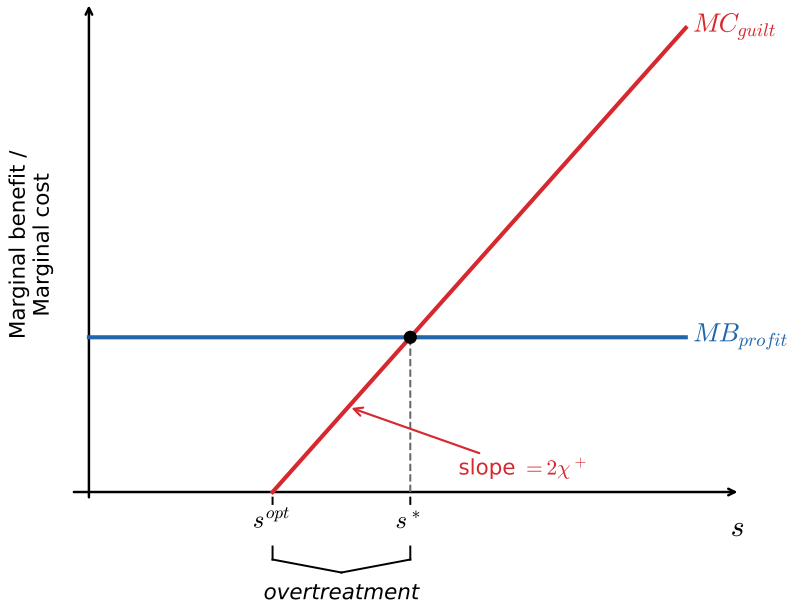


Figure 6: How χ^+ governs overtreatment

Notes: The marginal guilt cost of deviating from s^{opt} rises linearly with slope $2\chi^+$. Given a constant marginal profit incentive MB_{profit}, the provider chooses s^* where $MC_{\text{guilt}} = MB_{\text{profit}}$, so the overtreatment gap is $s^* - s^{\text{opt}} = MB_{\text{profit}}/(2\chi^+)$.

6 Estimation Results

6.1 Parameter estimates

Table 4 reports the estimated structural parameters. We impose symmetry in the discrete switching penalties: providers face a common penalty δ_{up} for upcoding (choosing $k > k^{\text{opt}}$) and δ_{down} for downcoding ($k < k^{\text{opt}}$), so that $\delta_{2,1} = \delta_{3,2} = \delta_{\text{up}}$ and $\delta_{1,2} = \delta_{2,3} = \delta_{\text{down}}$.

The estimated upcoding penalty ($\hat{\delta}_{\text{up}} = 0.274$) exceeds the downcoding penalty ($\hat{\delta}_{\text{down}} = 0.221$), indicating that providers face a larger non-pecuniary cost from choosing a more intensive procedure type than clinically warranted than from choosing a less intensive one.

We summarize the estimated distributions of χ^+ and χ^- by reporting the implied distribution of deviations from optimal treatment in response to a hypothetical constant marginal incentive of 0.1 points per one reimbursement unit (Panel D of Table 4). Figure 6 illustrates the calculation: given a constant marginal profit incentive MB_{profit}, the provider overtreats until the rising marginal guilt cost $MC_{\text{guilt}} = (s - s^{\text{opt}}) \cdot 2\chi^+$ equals that incentive, yielding chosen intensity s^*. A higher χ^+ steepens the MC_{guilt} line and therefore shrinks the overtreatment gap $s^* - s^{\text{opt}}$; conversely, a provider with low χ^+ deviates further. The same logic applies symmetrically for undertreatment governed by χ^-.

Table 4: Structural parameter estimates

Parameter	Description	Estimate
Panel A. Discrete switching penalties		
δ_{up}	Upcoding penalty ($k > k^{\text{opt}}$)	0.274
δ_{down}	Downcoding penalty ($k < k^{\text{opt}}$)	0.221
Panel B. Logit scale		
σ_ε	Type-I EV shock scale	0.438
Panel C. Provider heterogeneity (transformed beta)		
a_1	Shape parameter for χ^+ (overtreatment)	0.656
b_1	Shape parameter for χ^+ (overtreatment)	0.132
a_2	Shape parameter for χ^- (undertreatment)	10.245
b_2	Shape parameter for χ^- (undertreatment)	3.420
Panel D. Implied deviations from optimal treatment		
	Percentile	$s^* - s^{\text{opt}}$ (points)
Overtreatment (χ^+)		
	5th	31,594
	10th	8,111
	25th	119
	50th	≈ 0
Undertreatment (χ^-)		
	5th	2,180
	10th	1,494
	25th	694
	50th	211
	75th	33
	90th	2

Notes: Estimates obtained from the cross-regime deconvolution estimator described in Section 5. The switching penalties are restricted to depend only on the direction of deviation: $\delta_{2,1} = \delta_{3,2} \equiv \delta_{\text{up}}$ and $\delta_{1,2} = \delta_{2,3} \equiv \delta_{\text{down}}$. Provider heterogeneity parameters (a_1, b_1) and (a_2, b_2) govern beta distributions that, after the monotone transformation $g(\cdot)$ in Section A.1, yield the distributions of χ^+ and χ^- in the provider population. Panel D reports the implied distribution of treatment deviations $|s^* - s^{\text{opt}}|$ (in medical points) across providers in response to a hypothetical constant marginal profit incentive of 0.1 points per reimbursement unit, computed as $|s^* - s^{\text{opt}}| = 0.1/(2\chi) \times 10^5$ where $\chi = g(F^{-1}_{\text{Beta}}(q))$ at each percentile q. The factor of 10^5 reflects the rescaling of s to units of 10^5 points in the estimation; see Figure 6 for an illustration.

Table 5: Model fit: procedure-type shares

	Estimated			Data		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
2009	0.023	0.457	0.520	0.022	0.466	0.511
2010	0.023	0.421	0.556	0.026	0.427	0.548
2011	0.023	0.442	0.535	0.019	0.424	0.557

Notes: "Estimated" shares are the model-predicted procedure-type probabilities at the parameter estimates in Table 4, averaged over the estimated distribution of medical need. "Data" shares are the observed fractions in each year.

6.2 Model fit

We assess model fit along two dimensions: the discrete procedure-type shares and the full distribution of continuous spending intensity.

Procedure-type shares Table 5 compares the estimated and observed procedure-type shares by year. The model reproduces the data shares closely. In particular, the estimator captures the small but nonzero share of $k = 1$ (partial hip, low equipment) admissions and the shift from $k = 2$ toward $k = 3$ between 2009 and 2010, when the DRG reform changed the relative profitability of the two procedure types. In the data, the total hip share rises from 51.1% to 54.8% and the high-equipment partial hip share falls from 46.6% to 42.7%; the model tracks both movements (52.0% to 55.6% and 45.7% to 42.1%, respectively).

Spending distributions Figure 7 overlays the model-simulated distribution of medical points (orange) on the data distribution (blue). Each row is a year; columns show total hip and partial hip replacements. The green vertical line marks the DRG lower-bound threshold.

The model fits the data well across all years and procedure types. In 2009, before the DRG reform applies to partial hip replacements, the spending distributions are smooth and unimodal for both procedure types; the model reproduces the location, spread, and right skewness of these pre-reform distributions. In 2010, the first year of DRG reimbursement for partial hip replacements, both the data and the model show a sharp concentration of mass just above the lower-bound threshold—the bunching pattern documented in Section 3.4. For total hip replacements, where DRG reimbursement was already in place in 2009, the model captures the leftward shift in spending and the emergence of a more pronounced spike at the threshold once the lower bound is recalculated in 2010. In 2011, the equipment component Q falls, so the lower-bound threshold corresponds to a higher level of continuous spending s. This shifts the bunching point into the interior of the 2010 distribution, producing a distinct pattern that the model also tracks. For partial hip replacements, the model similarly captures the broader dispersion in spending and the year-to-year changes in the location of bunching as the threshold moves.

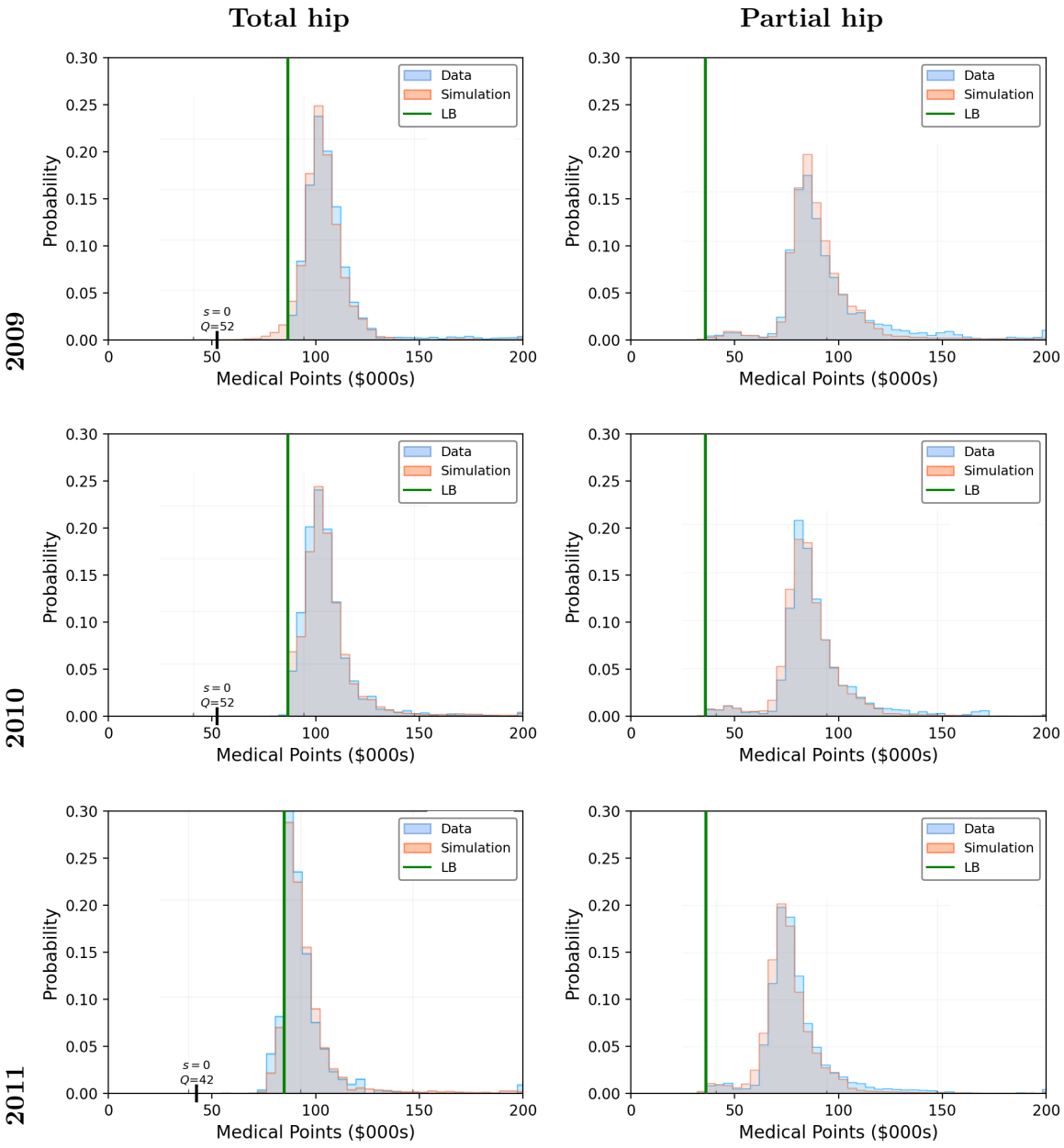


Figure 7: Model fit: spending distributions

Notes: Each row is a year (2009–2011); columns show total hip replacement ($k = 3$) and partial hip replacement ($k = 1, 2$). Blue: data; orange: simulation; green vertical line: DRG lower-bound threshold. Total reimbursement (the x-axis) is the sum of the equipment component $Q_P(k)$ and continuous spending s; these are marked on the total hip panels. In 2009 (pre-reform), the lower bound is not binding. In 2010, the lower bound binds and generates bunching just above the threshold. In 2011, Q falls, so the lower bound corresponds to a higher value of s falling in the interior of the 2010 distribution, producing a distinct bunching pattern. The model fits these year-specific patterns.

7 Simulations

7.1 Quantifying Treatment Distortions

With estimated preferences and the recovered distribution of medical need in hand, we can use the model to decompose observed treatment choices into clinically appropriate and incentive-driven components. We simulate 150,000 admissions per policy-year, drawing medical need from the estimated distribution and provider preferences from the estimated heterogeneity distribution. For each simulated admission, the model produces both the clinically optimal treatment $m^{\text{opt}} = (k^{\text{opt}}, s^{\text{opt}})$ and the chosen treatment (k, s). Table 6 compares resource costs under clinically optimal treatment with those under the model-predicted chosen treatment, broken down by procedure type. Equipment costs are held fixed at $c \cdot Q_{2011}(k)$ across all years so that cross-year comparisons isolate behavioral distortions from schedule changes in the equipment component. We begin with the discrete margin.

Table 6: Resource costs under clinically optimal vs. chosen treatment

	Optimal	2009	2010	2011
$k = 1$ *(partial, low equip.)*				
Fraction	0.023	0.023	0.023	0.023
Equipment cost	6,689	6,689	6,689	6,689
Mean continuous cost $c \cdot s$	39,099	46,985	40,504	40,781
$k = 2$ *(partial, high equip.)*				
Fraction	0.457	0.457	0.421	0.442
Equipment cost	28,633	28,633	28,633	28,633
Mean continuous cost $c \cdot s$	34,433	42,309	36,681	36,214
$k = 3$ *(total)*				
Fraction	0.520	0.520	0.556	0.535
Equipment cost	31,885	31,885	31,885	31,885
Mean continuous cost $c \cdot s$	43,584	41,642	45,169	45,861
Total resource cost	69,122	71,890	71,420	71,352

Notes: "Optimal" assigns each patient the clinically optimal treatment m^{opt}. Columns 2009–2011 report model-predicted treatment under the actual reimbursement regime. Equipment costs use $c \cdot Q_{2011}(k)$ in all years so that cross-year comparisons isolate behavioral distortions from schedule changes. Total resource cost aggregates equipment and continuous costs using the procedure-type fractions as weights. Based on 150,000 simulated admissions per year.

In 2009, the procedure-type fractions under the actual reimbursement regime are virtually identical to their clinically optimal counterparts: the k=2 share is 45.7% under both, and the k=3 share is 52.0% under both. The pre-reform regime—fee-for-service for partial hip replacements and DRG with no binding lower-bound threshold for total hip—provided little financial incentive to distort the discrete choice, and the model confirms that providers chose procedure types in line with clinical need. After the 2010 reform, the k=3 (total hip) share rises to 55.6%, overshooting the 52.0% optimum, while the k=2 (partial hip, high equipment) share falls to 42.1%, below its optimal 45.7%. This upcoding from partial to total hip is consistent with the DRG incentive to classify patients into

the higher-paying category. By 2011 the gap narrows but does not close: the k=3 share remains 1.5 percentage points above the optimum.

This upcoding is qualitatively different from distortions on the continuous margin. Continuous over- or undertreatment adjusts the intensity of a procedure that is itself appropriate for the patient; upcoding means performing a more invasive surgery than the patient's condition warrants. In this setting, patients who would optimally receive a partial hip replacement instead receive a total hip replacement—a procedure with greater surgical risk, longer recovery, and higher resource cost—purely because the reimbursement schedule makes the more invasive category more profitable.

Turning to continuous spending, the pre-reform pattern is one of substantial overtreatment within both partial hip categories. Mean continuous costs for k=1 and k=2 admissions exceed their clinically optimal levels by roughly 7,900 points in 2009, reflecting the positive marginal return to treatment intensity under fee-for-service. The DRG reform sharply reduces this overspending: by 2010, mean continuous costs for k=1 and k=2 fall to within 1,400 and 2,250 points of the optimum, respectively. For k=3 (total hip), where DRG reimbursement was already in place, the pattern is reversed: continuous spending is slightly below the optimum in 2009 and slightly above it in 2010–2011, consistent with the introduction of a binding lower-bound threshold that encourages bunching above the floor.

In aggregate, total resource costs fall from 71,890 in 2009 to 71,420 in 2010 and 71,352 in 2011, moving toward the clinically optimal level of 69,122 (approximately US$2,300 per admission at the prevailing exchange rate). The remaining excess of roughly 2,200–2,800 points (3–4%) reflects the offsetting effects of the reform: the large reduction in continuous overspending on partial hip replacements is partially offset by the shift toward more costly total hip procedures and modestly higher continuous spending within the total hip category.

Because Table 6 reports averages within each procedure type, a mean close to the optimum is consistent with either most patients receiving near-optimal treatment or some providers overtreating and others undertreating by similar amounts. Table 7 addresses this by reporting the share of individual admissions in which chosen treatment deviates from the patient-specific optimum. The admission-level patterns reinforce the average comparisons in Table 6: in 2009, overtreatment and undertreatment on the continuous margin are nearly balanced (48% vs. 52% of admissions), consistent with the modest average excess. After the reform, the distribution becomes sharply one-sided—93% of admissions involve $s < s^{\text{opt}}$ in 2010—so the lower average spending in the partial hip categories reflects a broad shift rather than a small number of large reductions. On the discrete margin, distortion rates are near zero in 2009, consistent with the close match between procedure-type shares and the optimum in Table 6. Upcoding rises to 3.5% of admissions in 2010, again consistent with the shift in fractions.

Table 7: Decomposition of treatment distortions (% of admissions)

	Continuous (s vs. s^{opt})		Discrete (k vs. k^{opt})	
	$s > s^{\text{opt}}$	$s < s^{\text{opt}}$	$k > k^{\text{opt}}$	$k < k^{\text{opt}}$
2009	48.00	52.00	0.04	0.07
2010	7.13	92.87	3.54	0.01
2011	13.24	86.76	1.63	0.13

Notes: Each entry is the percentage of simulated admissions in which the chosen treatment deviates from the clinically optimal treatment. Continuous distortions compare chosen intensity s to optimal intensity s^{opt}; discrete distortions compare chosen procedure type k to optimal type k^{opt}. Based on 150,000 simulated admissions per year.

8 Reimbursement Schedule Design

Section 7.1 showed that the observed schedules trade off reductions in continuous overspending against increases in upcoding and threshold bunching. We now use the estimated model to seek reimbursement schedule that best navigates this trade-off for a fixed budget. Throughout, we use "optimal" as shorthand for distortion-minimizing relative to a specific planner loss and, where noted, relative to a restricted schedule class; we do not claim welfare optimality in a broader sense.

Formally, we consider a planner who chooses a schedule $W(k, s)$ to minimize a loss function $\mathcal{L}(W)$ measuring treatment distortion, subject to the budget-balance constraint $\mathbb{E}[W(k^*, s^*)] = \bar{B}$, where $\bar{B}$ is the expected reimbursement under the 2011 regime and the expectation is taken at the equilibrium treatment choices (k^*, s^*) induced by W. The remainder of the section develops this exercise in three steps. Section 8.1 shows that the unconstrained problem admits a trivial closed-form solution—known *ex ante*, before any estimation—which motivates restricting attention to institutionally realistic schedules. Section 8.2 formulates the distortion-minimizing problem within the TW-DRG functional form, specifies the planner's loss, and sketches the computational approach; full algorithmic detail is deferred to Appendix D.

Discussion For a given budget, two constraints need to be satisfied to implment a payment schedule: (1) participation and (2) incentive compatibility. With information on m^{opt}, the planner could design the schedule to satisfy both constraints *at the case level*, (e.g. Gaynor et al., 2023b). Absent of this information, such optimal contracts are unlikely to be feasible. Our counterfactual experiments satisfy (2) at the case level by solving for providers' optimal choices and (1) in expectation.

8.1 Marginal-cost reimbursement

In our framework, doctors' incentives can be aligned with clinically optimal treatment by reimbursing exactly at marginal cost and holding profit fixed via a lump-sum transfer. Concretely, consider the schedule

$$W^{FB}(k, s) \;=\; \alpha + \underbrace{C_0(k) + c\,s}_{\text{reimbursed at marginal cost}}, \tag{21}$$

where $C_0(k)$ is the fixed resource cost of procedure type k and $c \in (0,1)$ is the marginal resource cost of a unit of billed intensity (both from (4)); the lump-sum constant α is chosen to satisfy the budget constraint. Under W^{FB}, profit as defined in (5) reduces to

$$\pi(k,s) \;=\; W^{FB}(k,s) - \underbrace{\big(C_0(k) + c\,s\big)}_{\text{resource cost}} \;=\; \alpha$$

for every treatment bundle (k,s). Two consequences follow. On the continuous margin, marginal profit in s is zero, so the provider's first-order condition reduces to $\partial\gamma_s/\partial s = 0$, which by (7) is solved at $s^* = s^{\text{opt}}$ for every realization of χ. On the discrete margin, profit does not vary with k, so the switching penalties $\delta_{k,k^{\text{opt}}} \geq 0$ push each provider toward $k^* = k^{\text{opt}}$; the only residual variation across simulated admissions comes from the logit shocks ε_k.

To see why this aligns incentives with optimal treatment, recall that in the model, the only latent object is the patient-specific target m^{opt}, and the guilt term in (6) pushes the provider toward m^{opt} whenever profit does not pull her away. Because the payer observes the realized treatment (k,s) and can price every observable action separately, the instrument space is rich enough to set marginal profit to zero on both margins simultaneously—and a flat-profit schedule is all that is required to deactivate the financial incentive to deviate. The first-best schedule is, in this sense, a pure cost-reimbursement contract plus a lump-sum transfer: it pays the provider exactly her resource cost on the margin, eliminating the moral-hazard channel entirely.

It is worth emphasizing that W^{FB} is *not* the 2009 pre-reform schedule: for partial hip replacements, marginal reimbursement in s equaled one in 2009, so marginal profit was $1 - c > 0$ and the doctor overtreated. The first-best instead reimburses intensity at rate c, which exactly offsets the marginal resource cost.

8.2 Optimization within the DRG structure

While the distortion-minimizing schedule is trivial and known *ex ante*—before estimating either the guilt parameters or the distribution of clinically optimal need—policymakers typically pursue additional goals that amount to optimizing a different objective or imposing constraints on the form of the schedule. Taiwan's switch to a prospective DRG system, for example, was motivated in part by a desire to reduce administrative costs and contain aggregate spending growth, considerations that the loss function $\mathcal{L}$ does not price. In that spirit, this section asks what the distortion-minimizing schedule looks like *within* the DRG functional form: what is the best schedule a planner could choose if she had to deliver care under a TW-style DRG structure?

This exercise also connects directly to Section 7.1: there we measured how the observed DRG redistributes treatment across the three distortion margins relative to fee-for-service; here we ask how much further that redistribution could be pushed by re-choosing the parameters of the same institutional form.

Recall that the TW-DRG schedule pays a fee-for-service rate below a lower spending threshold, a fixed prospective payment in an interior flat region, and a fee-for-service "outlier" rate above an

upper threshold:

$$W(k,s) = \begin{cases} s + Q(k), & s + Q(k) < s^{LB}(k), \\ s^f(k), & s^{LB}(k) \le s + Q(k) < s^{UB}(k), \\ s^f(k) + r\left(s + Q(k) - s^{UB}(k)\right), & s + Q(k) \ge s^{UB}(k), \end{cases} \tag{22}$$

where $Q(k)$ is the procedure-specific equipment add-on. In its most general form, this functional form has as free parameters the two thresholds $s^{LB}(k)$ and $s^{UB}(k)$, the flat payment $s^f(k)$, and the equipment add-on $Q(k)$ for each procedure type k, together with all three marginal rates (the FFS rate below, the flat rate in the interior, and the outlier rate r). A preliminary observation is that optimizing over all of these parameters does not yield a substantive problem: setting both the below-threshold rate and the outlier rate to c, collapsing the flat region to a point, and using $s^f(k)$ as a lump-sum shifter reproduces the first-best (21) exactly.

We therefore fix the defining institutional features of the DRG system—the marginal rates $\{1, 0, 0.8\}$ and the equipment add-ons $Q(k)$ at their 2011 values—and let the planner optimize over the threshold locations and flat-payment levels. As in the actual 2011 schedule, the two partial hip categories $k = 1$ and $k = 2$ share a common set of DRG parameters, and total hip ($k = 3$) has its own, giving six free parameters:

$$\phi = \left\{ s^{LB}_{\text{partial}},\ s^{UB}_{\text{partial}},\ s^{f}_{\text{partial}},\ s^{LB}_{\text{total}},\ s^{UB}_{\text{total}},\ s^{f}_{\text{total}} \right\}.$$

The planner's problem is to choose ϕ to minimize the distortion loss $\mathcal{L}(W(\cdot;\phi))$ subject to two sets of constraints: the budget-balance constraint $\mathbb{E}[W(k^*, s^*)] = \bar{B}$, and a pair of *coverage constraints* fixing the equilibrium share of admissions in the flat region at target values X_{partial} and X_{total}, separately by procedure group $g \in \{\text{partial}, \text{total}\}$:

$$\Pr\left(s^{LB}_g \le s^* + Q(k) < s^{UB}_g \,\middle|\, k \in g;\, \phi\right) \;=\; X_g, \tag{23}$$

where the probability is taken over the equilibrium distribution of s^* induced by ϕ. Without the coverage constraints, nothing prevents the optimizer from pushing the flat region out to the edge of the spending distribution, effectively removing the prospective-payment feature and producing a schedule that is "DRG in name only." We report results along a loss-coverage frontier, varying $X = (X_{\text{partial}}, X_{\text{total}})$ via a common scalar multiplier on the two observed 2011 coverage rates; the actual 2011 values provide a natural reference point. In addition, we impose the ordering constraint $s^{LB}_g < s^{UB}_g$ and a non-negative-notch-size condition $s^f_g \ge s^{LB}_g$, which rules out a downward jump at the lower threshold (the schedule may still jump up, producing the usual notch).

Trade-offs introduced by the restriction. Holding the marginal rates fixed means no region reimburses at rate c, so profit is never flat in s: the flat region creates undertreatment incentives, the FFS tail creates overtreatment incentives, and the notch at s^{LB} creates bunching. The planner

trades off these three distortions by choosing where to place the thresholds and how high to set the flat payment.

Why the problem is not standard. Three features make the planner's problem nontrivial. First, the equilibrium spending distribution is itself endogenous to ϕ: each candidate schedule induces a different distribution of optimal choices (k^*, s^*), so both the budget moment $\mathbb{E}[W(k^*, s^*)]$ and the coverage probability in (23) must be re-evaluated by re-solving the provider's problem at every iteration. Second, the coverage constraint is a fixed-point in ϕ: the thresholds $s^{LB}(k)$ and $s^{UB}(k)$ that bound the flat region also help determine—through the bunching they induce at the lower notch and the FFS regions they expose—the very distribution of s^* whose mass in that region is being constrained. Third, the lower-notch bunching produces a mass point in the equilibrium distribution of $s^* + Q(k)$, so the coverage probability is a step function of ϕ wherever the bunching point crosses a threshold; the loss function inherits this non-smoothness.

Loss function. We measure treatment distortion using the squared resource-cost gap between the realized and the clinically optimal treatment bundle. Let $C(k, s) \equiv C_0(k) + c\, s$ denote the resource cost of bundle (k, s), as in (4). The planner's loss is

$$\mathcal{L}(W) \;=\; \mathbb{E}\Big[\big(C(k^*, s^*) - C(k^{\text{opt}}, s^{\text{opt}})\big)^2\Big], \tag{24}$$

where the expectation is taken under the equilibrium distribution of (k^*, s^*) induced by W. Two features of this loss are worth emphasizing. First, expressing both margins in resource-cost units puts intensity deviations and procedure misallocations on a common footing without introducing a free weighting parameter: a one-unit deviation in s contributes c^2 to the loss, while substituting procedure k for k^{opt} contributes $(C_0(k) - C_0(k^{\text{opt}}))^2$ plus a cross term in the intensity deviation. Second, the resource-cost gap has a direct normative interpretation in our framework: by (4), $C(k, s)$ is precisely the quantity whose minimization, holding $(k^{\text{opt}}, s^{\text{opt}})$ fixed at the patient's clinical target, constitutes appropriate care. The loss $\mathcal{L}$ penalizes departures from that target symmetrically in either direction, scaled by the real-resource consequences of the departure.

This objective compresses all three distortion margins studied in Section 7.1 into a single scalar. To keep the connection to those margins transparent, we report companion outcomes alongside $\mathcal{L}$ when we evaluate schedules: the upcoding rate $\Pr(k^* \neq k^{\text{opt}})$, the overtreatment share $\Pr(s^* > s^{\text{opt}})$, the undertreatment share $\Pr(s^* < s^{\text{opt}})$, the mean absolute intensity deviation $\mathbb{E}|s^* - s^{\text{opt}}|$, and the flat-region coverage rates. These companion outcomes let the reader see how the distortion-minimizing schedule rebalances the three margins relative to the 2011 DRG and 2009 pre-reform benchmarks.

Computational approach. Solving the problem requires re-simulating equilibrium at each candidate ϕ and handling the fixed-point in the coverage constraints. We hold fixed a simulation panel of patient-provider draws $\{(m_i^{\text{opt}}, \chi_i, \varepsilon_i)\}_{i=1}^N$ from the estimated population distribution, and at each

trial ϕ re-solve the provider problem of Section 4.3 under $W(\cdot;\phi)$ to obtain sample analogs of the budget moment, the coverage probabilities, and the loss. The coverage constraints are imposed by nesting a one-dimensional root-find on each group's upper threshold s_g^{UB} inside an outer sequential quadratic programming step over the remaining four parameters, with the budget constraint enforced as an equality at the outer level. We smooth through the notch-induced mass point during the inner search and verify the un-smoothed constraints at convergence. Appendix D gives the full algorithm.

8.3 Results

9 Conclusion

This paper develops and estimates a structural model of physicians' treatment decisions under alternative reimbursement systems, using data on hip replacement surgeries in Taiwan during a period of major payment reform. The model jointly recovers the population distribution of clinically optimal treatment and the distribution of physicians' nonpecuniary costs of deviating from it, using cross-regime variation in reimbursement schedules to separately identify three margins of distortion: overtreatment at the intensive margin, undertreatment at the intensive margin, and upcoding across procedure categories. The estimated model fits the key patterns in the data, including the shift in procedure-type shares after the reform, the reduction in average spending under DRG reimbursement, and the emergence of bunching at DRG payment thresholds. We are currently using the estimated model to study the optimal design of reimbursement schedules that minimize deviations from clinically optimal treatment.

References

Alexander, D. and Schnell, M. (2024). The impacts of physician payments on patient access, use, and health. *American Economic Journal: Applied Economics*, 16(3):142–177.

Barnett, M. L., Wilcock, A., McWilliams, J. M., Epstein, A. M., Joynt Maddox, K. E., Orav, E. J., Grabowski, D. C., and Mehrotra, A. (2019). Two-year evaluation of mandatory bundled payments for joint replacement. *The New England Journal of Medicine*, 380(3):252–262.

Barros, P. and Braun, G. (2017). Upcoding in a national health service: The evidence from Portugal. *Health Economics*, 26(5):600–618.

Bertanha, M., McCallum, A. H., and Seegert, N. (2023). Better bunching, nicer notching. *Journal of Econometrics*, 237(2):105427.

Blomquist, S., Newey, W. K., Kumar, A., and Liang, C.-Y. (2021). On bunching and identification of the taxable income elasticity. *Journal of Political Economy*, 129(8):2320–2343.

Carter, G. M., Newhouse, J. P., and Relles, D. A. (1990). How much change in the case mix index is DRG creep? *Journal of Health Economics*, 9(4):411–428.

Chan, D. C., Gentzkow, M., and Yu, C. (2022). Selection with variation in diagnostic skill: Evidence from radiologists. *Quarterly Journal of Economics*, 137(2):1173–1224.

Chen, P.-C., Kuo, N.-C., Huang, Y.-T., Huang, K.-H., Yang, M.-C., and Lee, Y.-C. (2016). Impact of DRG-based payments on health care providers. *Taiwan Journal of Public Health*, 35(3):268–280.

Cheng, S.-H., Chen, C.-C., and Tsai, S.-L. (2012). The impacts of DRG-based payments on health care provider behaviors under a universal coverage system: A population-based study. *Health Policy*, 107(2-3):202–208.

Chetty, R., Friedman, J. N., Olsen, T., and Pistaferri, L. (2011). Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: Evidence from danish tax records. *The Quarterly Journal of Economics*, 126(2):749–804.

Chien, L.-C., Chou, Y.-J., Huang, Y.-C., Shen, Y.-J., and Huang, N. (2020). Reducing low value services in surgical inpatients in Taiwan: Does diagnosis-related group payment work? *Health Policy*, 124(1):89–96.

Clemens, J. and Gottlieb, J. D. (2014). Do physicians' financial incentives affect medical treatment and patient health? *American Economic Review*, 104(4):1320–1349.

Coulam, R. F. and Gaumer, G. L. (1992). Medicare's prospective payment system: A critical appraisal. *Health Care Financing Review*, 1991(Suppl):45–77.

Cutler, D. M. (1995). The incidence of adverse medical outcomes under prospective payment. *Econometrica*, 63(1):29–50.

Dafny, L. S. (2005). How do hospitals respond to price changes? *American Economic Review*, 95(5):1525–1547.

Dranove, D. (1988). Demand inducement and the physician/patient relationship. *Economic Inquiry*, 26(2):281–298.

Dulleck, U. and Kerschbamer, R. (2006). On doctors, mechanics, and computer specialists: The economics of credence goods. *Journal of Economic Literature*, 44(1):5–42.

Dummit, L. A., Kahvecioglu, D., Marrufo, G., Rajkumar, R., Marshall, J., Tan, E., Press, M. J., Flood, S., Muldoon, L. D., Gu, Q., Hassol, A., Bott, D. M., Bassano, A., and Conway, P. H. (2016). Association between hospital participation in a Medicare bundled payment initiative and payments and quality outcomes for lower extremity joint replacement episodes. *JAMA*, 316(12):1267–1278.

Einav, L., Finkelstein, A., and Mahoney, N. (2018). Provider incentives and healthcare costs: Evidence from long-term care hospitals. *Econometrica*, 86(6):2161–2219.

Eliason, P. J., Grieco, P. L. E., McDevitt, R. C., and Roberts, J. W. (2018). Strategic patient discharge: The case of long-term care hospitals. *American Economic Review*, 108(11):3232–3265.

Ellis, R. P. and McGuire, T. G. (1986). Provider behavior under prospective reimbursement: Cost sharing and supply. *Journal of Health Economics*, 5(2):129–151.

Finkelstein, A., Ji, Y., Mahoney, N., and Skinner, J. (2018). Mandatory Medicare bundled payment program for lower extremity joint replacement and discharge to institutional postacute care: Interim analysis of the first year of a 5-year randomized trial. *JAMA*, 320(9):892–900.

Gaynor, M., Mehta, N., and Richards-Shubik, S. (2023a). Optimal contracting with altruistic agents: Medicare payments for dialysis drugs. *American Economic Review*, 113(6):1530–1571.

Gaynor, M., Mehta, N., and Richards-Shubik, S. (2023b). Optimal contracting with altruistic agents: Medicare payments for dialysis drugs. *American Economic Review*.

Geruso, M. and Layton, T. (2020). Upcoding: Evidence from Medicare on squishy risk adjustment. *Journal of Political Economy*, 128(3):984–1026.

Gowrisankaran, G., Joiner, K. A., and Lin, J. (2023). Costs, benefits, and the efficiency of hospital coding under Medicare. *American Economic Review*, 113(7):1879–1913.

Groß, M., Jürges, H., and Wiesen, D. (2021). The effects of audits and fines on upcoding in neonatology. *Health Economics*, 30(8):1978–1986.

Gruber, J., Kim, J., and Mayzlin, D. (1999). Physician fees and procedure intensity: the case of cesarean delivery. *Journal of Health Economics*, 18(4):473–490.

Gupta, A. (2021). Impacts of performance pay for hospitals: The readmissions reduction program. *American Economic Review*, 111(4):1241–1283.

Handel, B. R., Heizlsperger, L.-J., Knecht, J., Kolstad, J. T., Malmendier, U., and Matějka, F. (2026). Thinking versus doing: Cognitive capacity, decision making and medical diagnosis. Technical report, National Bureau of Economic Research.

Hodgkin, D. and McGuire, T. G. (1994). Payment levels and hospital response to prospective payment. *Journal of Health Economics*, 13(1):1–29.

Hu, W.-Y., Yeh, C.-F., Shiao, A.-S., and Tu, T.-Y. (2015). Effects of diagnosis-related group payment on health-care provider behaviors: A consecutive three-period study. *Journal of the Chinese Medical Association*, 78(11):678–685.

Jack, W. (2005). Purchasing health care services from providers with unknown altruism. *Journal of Health Economics*, 24(1):73 – 93.

Jürges, H. and Köberlein, J. (2015). What explains DRG upcoding in neonatology? the roles of financial incentives and infant health. *Journal of Health Economics*, 43:13–26.

Kleven, H. J. (2016). Bunching. *Annual Review of Economics*, 8(1):435–464.

Kleven, H. J. and Waseem, M. (2013). Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from Pakistan. *Quarterly Journal of Economics*, 128(2):669–723.

Lo, Y.-T., Chang, G.-M., and Tung, Y.-C. (2015). Impact of the DRG-based prospective payment system on medical utilization and outcomes of care for hip fracture and hip replacement. *Taiwan Journal of Public Health*, 34(2):180–192.

McGuire, T. G. and Pauly, M. V. (1991). Physician response to fee changes with multiple payers. *Journal of Health Economics*, 10(4):385–410.

Mullainathan, S. and Obermeyer, Z. (2022). Diagnosing physician error: A machine learning approach to low-value health care. *The Quarterly Journal of Economics*, 137(2):679–727.

Navathe, A. S., Liao, J. M., Dykstra, S. E., Wang, E., Lyon, Z. M., Shah, Y., Martinez, J., Small, D. S., Werner, R. M., Dinh, C., Ma, X., and Emanuel, E. J. (2018). Association of hospital participation in a Medicare bundled payment program with volume and case mix of lower extremity joint replacement episodes. *JAMA*, 320(9):901–910.

Newhouse, J. P. and Byrne, D. J. (1988). Did Medicare's prospective payment system cause length of stay to fall? *Journal of Health Economics*, 7(4):413–416.

Parkinson, B., Meacock, R., and Sutton, M. (2019). How do hospitals respond to price changes in emergency departments? *Health Economics*, 28(7):830–842.

Sacarny, A. (2018). Adoption and learning across hospitals: The case of a revenue-generating practice. *Journal of Health Economics*, 60:142–164.

Saez, E. (2010). Do taxpayers bunch at kink points? *American Economic Journal: Economic Policy*, 2(3):180–212.

Silverman, E. and Skinner, J. (2004). Medicare upcoding and hospital ownership. *Journal of Health Economics*, 23(2):369–389.

Simborg, D. W. (1981). DRG creep: A new hospital-acquired disease. *The New England Journal of Medicine*, 304(26):1602–1604.

Train, K. E. (2008). EM algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling*, 1(1):40–69.

Tung, Y.-C., Chang, H.-Y., and Chang, G.-M. (2018). Impact of bundled payments on hip fracture outcomes: A nationwide population-based study. *International Journal for Quality in Health Care*, 30(1):23–31.

Yan, Y.-H., Kung, C.-M., and Chen, Y. (2017). The exploration of medical resources utilization among inguinal hernia repair in Taiwan diagnosis-related groups. *BMC Health Services Research*, 17(1):708.

Appendix

A Detailed Functional Forms

A.1 Provider heterogeneity $G(\chi)$

Each of χ_i^+ and χ_i^- has support on $(0, \infty)$, and we need a flexible specification of its density. We use a beta density—which provides a rich two-parameter family of shapes on $(0, 1)$—and transform it to $(0, \infty)$ via a fixed monotone map. This gives each marginal distribution two shape parameters while keeping the construction tractable.

Specifically, let $g : (0, 1) \to (0, \infty)$ be defined by

$$\chi = g(z) \equiv \exp\Big(\frac{1}{1-z} - 1\Big) - 1, \qquad z = g^{-1}(\chi) \equiv 1 - \frac{1}{1 + \ln(1+\chi)}.$$

We draw latent $(z_{1,i}, z_{2,i}) \in (0, 1)^2$ with beta marginals,

$$z_{1,i} \sim \text{Beta}(a_1, b_1), \qquad z_{2,i} \sim \text{Beta}(a_2, b_2),$$

and convert to the curvature parameters by $\chi_i^+ = g(z_{1,i})$ and $\chi_i^- = g(z_{2,i})$. In the baseline implementation, we impose comonotonic dependence (perfect rank dependence) between $z_{1,i}$ and $z_{2,i}$ by drawing a single percentile $u_i \sim \text{Unif}(0, 1)$ and setting

$$z_{1,i} = F^{-1}_{\text{Beta}(a_1,b_1)}(u_i), \qquad z_{2,i} = F^{-1}_{\text{Beta}(a_2,b_2)}(u_i),$$

so that (χ_i^+, χ_i^-) inherit the same rank across providers while allowing different marginal shapes.