



XLV

Jornadas de

Economía de la Salud

Datos, evidencia, decisiones:
generando valor para la gestión
y las políticas sanitarias

Sevilla, 17 al 19 de junio de 2026

Who Benefits from Public vs. Private Hospital Care? Patient-Level Heterogeneity in Surgical Outcomes

Carlos Gallego-Moll^{1,2} · **Laia Maynou**^{1,3}

¹Universitat de Barcelona; ²Universidad Pública de Navarra;

³London School of Economics and Political Science

Introduction and overview

Motivation. Public systems increasingly contract elective surgery to private hospitals. Does ownership change **efficiency** or **quality**?

Questions. (1) Are there differences between being treated in a **public vs. private hospital**?
(2) **For whom**: do those differences vary across patient subgroups?

Methods. 99,567 elective procedures, 4 surgery types, 8 contrasts; three estimators: multilevel regression → PSM → ML (GBM + causal forest).

Findings. Only cataracts move: public stays **≈5–6% longer** than for-profit, stable across estimators. ML corrects one inflated classical estimate. **No** patient-level heterogeneity detected.

Why does hospital *ownership* matter?

- Public health systems in high-income countries face sustained fiscal pressure \Rightarrow they increasingly **contract elective surgery to private hospitals** while keeping financing public.
- Does ownership change the **quality** or **efficiency** of care delivered under that single public contract?
- Evidence is mixed and US/UK-dominated; capacity-referral systems are under-studied.
- Catalonia splits financing from provision: **$\approx 70\%$ of hospitals are privately owned**, yet most surgery is **publicly financed**.

Institutional context: how patients reach private hospitals

- Tax-funded, universal system (CatSalut). Public hospitals on global budgets; non-profit and for-profit hospitals paid **per case**.
- **Decret 354/2002** waiting-time guarantee: when a public hospital cannot meet the deadline, CatSalut **refers the patient to a contracted private hospital**.
- Private-sector patients are selected on **capacity constraints** ⇒ referral is administrative.
- For-profit hospitals specialise in **high-volume, low-complexity** elective episodes; public tertiary hospitals keep emergencies and complex cases.

Two empirical gaps in the public–private comparison

- **Selection / cream-skimming.** Private hospitals concentrate on simpler cases; raw differences mix case-mix and ownership.
- **Hidden heterogeneity.** Average effects can mask subgroups.
- **The gap we fill.** Prior work estimates ownership effects *on average*, in patient-choice systems (US, English NHS). We add two things: the average effect in a **capacity-referral** single-payer system, and whether that gap *varies across patients* — “for whom?”

Research questions

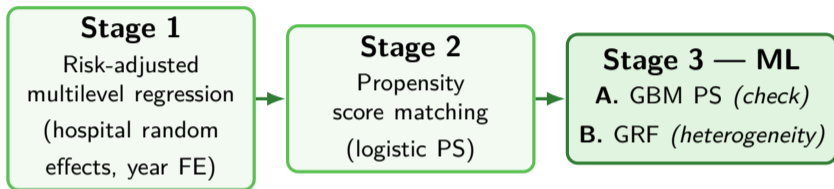
1. Do private hospitals deliver **shorter stays** or **lower readmission** than public, after risk adjustment?
2. How much of any gap **survives matching** and a flexible propensity model?
3. Does the public–private gap **vary across patient subgroups** (age, morbidity, SES, region)?

Data: PADRIS population registry, Catalonia 2011–2019

- 10% random sample, aged 60+
⇒ **N = 99,567** elective procedures.
- Outcomes: **Length of stay** (days) and **30-day readmission**.
- Two contrasts (public vs. for-profit / non-profit) × 4 procedures = **8 analyses**.
- Covariates: age bands, sex, SES (copayment), AMG morbidity, region, year.

Procedure	N
Cataract surgery	82,809
Hip & knee replacement	9,786
Prostatectomy	3,699
Hysterectomy	1,816
Total	99,567

Methods

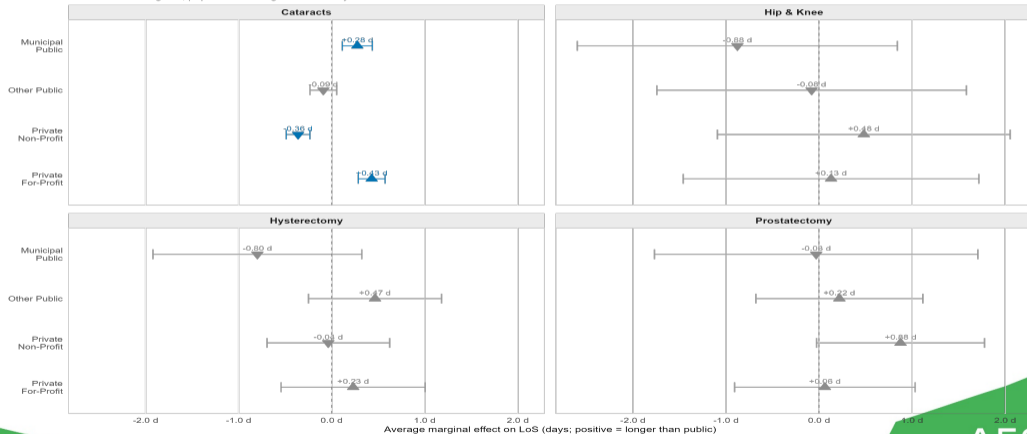


- Same identifying assumptions (**unconfoundedness**, overlap).
- Estimand: **ATT** (effect on the treated). $W = 1$ public \Rightarrow positive = longer public stays.
- **GBM (Module A)** is a **specification check** on the propensity score: it relaxes the logistic functional form and re-tests the ATT.
- **GRF causal forest (Module B)** targets **heterogeneity** (CATE): for which subgroups of patients the gap is larger or smaller.

Results: average length-of-stay effects — only cataracts moves

Length of Stay — Average Marginal Effect vs. Public Hospitals

Gamma GLMM log-link, population-average · AME in days · 95% CIs



▽ Lower than public △ Higher than public ● p ≥ 0.05 ● p < 0.05

△ = Higher than public ▽ = Lower than public ○ = No difference
 Blue = p < 0.05 Grey = p < 0.05 Reference line = 0 (no effect on days)

Results: ATT across estimators — flexible PS moderates the extreme values

Procedure	Comparison	Matching (PSM)		Forest
		Logit	GBM	GRF
Cataracts	Pub. vs. FP	+0.103***	+0.054***	+0.093
Cataracts	Pub. vs. NP	+0.017	+0.006	-0.013
Hip & Knee	Pub. vs. FP	-0.025	-0.005	-0.070
Hip & Knee	Pub. vs. NP	-0.020	-0.013	-0.004
Prostatectomy	Pub. vs. FP	-0.002	-0.055	-0.072
Prostatectomy	Pub. vs. NP	-0.888***	-0.165***	-0.071
Hysterectomy	Pub. vs. FP	-0.054	-0.026	+0.021
Hysterectomy	Pub. vs. NP	+0.271	+0.041	+0.018

LoS, log-days. * $p < 0.05$,

** $p < 0.01$, *** $p < 0.001$. ATT = Public – Private. GRF: doubly-robust AIPW. FP/NP = for-/non-profit.

Results: the cataract gap is robust across estimators



- Public hospitals keep cataract patients \approx **5–6% longer** than matched for-profit hospitals (log-days; GBM $p < 0.01$).
- All three estimators **agree in sign**; flexible propensity (GBM) **halves** the logistic estimate \Rightarrow it absorbs **non-linear confounding** the logistic model missed.
- Modest per patient (≈ 0.1 d on a 1.8-d baseline) but non-trivial at **>80,000** procedures.

Results: ML disciplines an implausible estimate

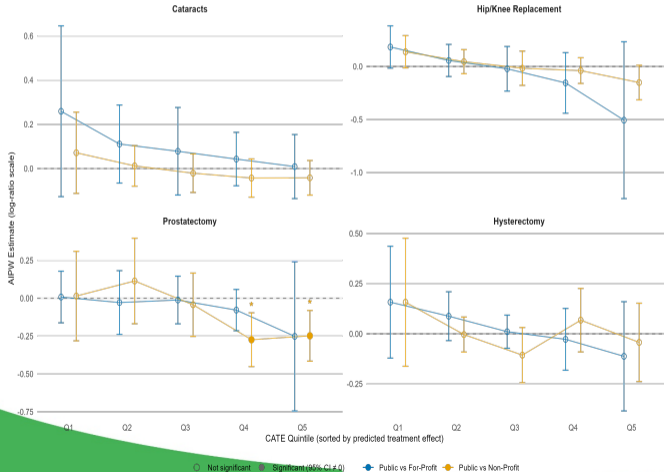


- Prostatectomy, public vs. non-profit: logistic PSM implies a **five-fold inflated** effect on a small, high-dimensional stratum.
- GBM and GRF bring it down to a moderate, **directionally consistent** effect.
- Flexible propensity estimation matters most where **matched samples are small**.

Results: no detectable patient-level heterogeneity

CATE Heterogeneity Profile by Quintile — All Four Procedures

Doubly-robust AIPW estimates of LoS effect (log scale) - 95% CIs - y-axis free per panel



- BLP calibration test non-significant in **all 8 contrasts** (BH $p > 0.999$), incl. no-region forests.
- CATE spread is **modest**; no distribution suggests clinically distinct subgroups.
- The forests **do not learn valid CATE rankings** (anti-calibrated) \Rightarrow quintile spreads are descriptive, not subgroup effects.

Results: 30-day readmission is underpowered

- Baseline prevalence only **1–7%** across procedures.
- Only cataract public–non-profit (AME +0.002, $p = 0.045$) is significant.
- Forest ATTs negligible ($|\text{ATT}| \leq 0.008$); BLP calibration fails in most contrasts.
- Null readmission findings reflect a **power limitation**.

Conclusions & policy implications

1. Public cataract stays are \approx **5–6% longer** than for-profit, robust across three estimators; small per patient, system-relevant at scale.
2. **No subgroup** demonstrably gains or loses more from public vs. private provision in these data.
3. **Equity reading.** A null is **compatible** with risk-adjusted capitation (MEDEA, AMG) and **capacity-based referral** that allocates patients to private hospitals on availability, not patient type.
4. Flexible ML propensity scores act as **specification checks**, correcting one implausible classical estimate.



Jornadas de
Economía de la Salud

Datos, evidencia, decisiones:
generando valor para la gestión
y las políticas sanitarias

Sevilla, 17 al 19 de junio de 2026

Thank you

Carlos Gallego-Moll^{1,2} · **Laia Maynou**^{1,3}

¹Universitat de Barcelona; ²Universidad Pública de Navarra; ³London School of Economics and Political Science

Funded by MICIN (Govt. of Spain) CNS2023-144351, ERDF/EU. Data: AQuAS / PADRIS.

A1. Identification & doubly-robust estimation

- Assumptions (all stages): conditional ignorability, overlap $0 < e(X) < 1$, SUTVA. ML relaxes functional form, **not** identification.
- ATT via the AIPW doubly-robust score (consistent if *either* \hat{e} or \hat{m}_0 is correct):

$$\hat{\tau}_{\text{ATT}} = \frac{1}{\sum_i W_i} \sum_i \left[W_i (Y_i - \hat{m}_0(X_i)) - (1 - W_i) \frac{\hat{e}(X_i)}{1 - \hat{e}(X_i)} (Y_i - \hat{m}_0(X_i)) \right]$$

- K -fold cross-fitting (GBM: out-of-bag; GRF: honest splitting) restores valid \sqrt{n} inference.

A2. GBM propensity scores & balance

- Gradient boosted trees ($\eta = 0.01$, depth 4); iteration count M^* chosen to **minimise covariate imbalance** (twang), *not* predictive accuracy.
- Rationale: high-dimensional AMG \times age \times SES \times region space; logistic PS unlikely to capture non-linearities; boosting reliable when treated fraction $< 10\%$.
- Balance (48 indicators, 8 contrasts): global **mean SMD = 0.014**, **99.4%** of pairs < 0.10 ; GBM at least as good as logistic everywhere.

A3. Causal forest & the BLP calibration test

- GRF: kernel-weighted residual-on-residual regression; $B = 2,000$ trees, honest splitting, **hospital cluster-robust** variance.

- Best linear predictor (calibration):

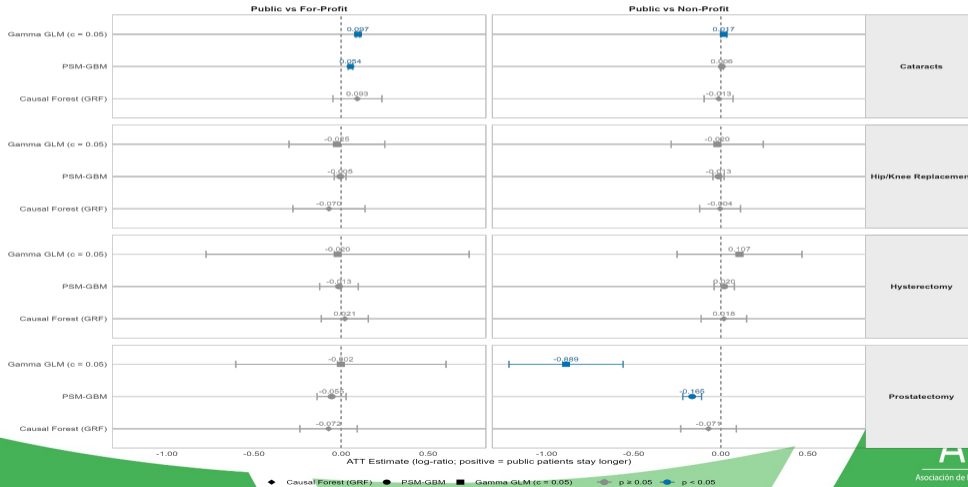
$$\tilde{Y}_i = a_0 \tilde{W}_i + a_1 (\hat{\tau}^{-i}(X_i) - \bar{\tau}) \tilde{W}_i + \varepsilon_i$$

- \hat{a}_0 (mean calibration) **passes**; \hat{a}_1 (heterogeneity) is **< 0 in all 8 contrasts** \Rightarrow rankings uninformative.
- Why ATT still valid: double robustness keeps the AIPW average consistent even when CATE rankings are noisy.

A4. GBM vs. logistic matching: length of stay

ATT Estimates for Length of Stay — Method Comparison

Gamma GLM (c = 0.05) · PSM-GBM · Causal Forest (GRF) · log-ratio scale



A5. Causal-forest summary & heterogeneity tests (LoS)

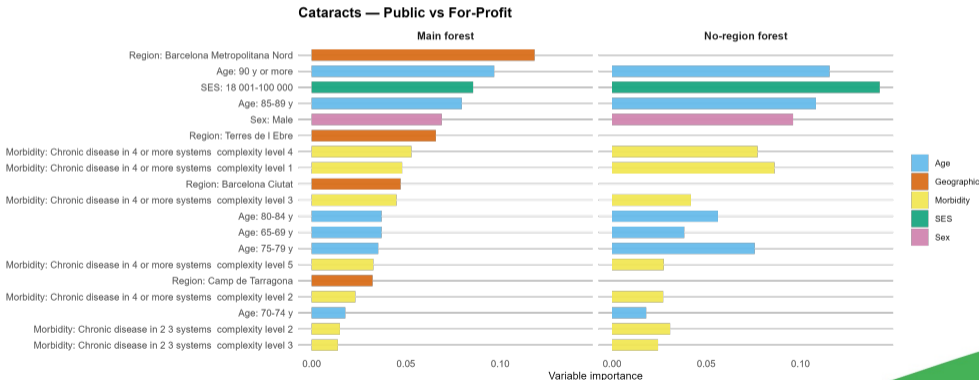
Proc.	Comp.	ATT	ATE _{ov}	Het. p	RATE p	No-reg. p
Cataracts	FP	+0.093	+0.051	>0.999	0.339	>0.999
Cataracts	NP	-0.013	+0.001	>0.999	0.339	>0.999
Hip & Knee	FP	-0.070	-0.019	>0.999	0.278	>0.999
Hip & Knee	NP	-0.004	-0.000	>0.999	0.158	>0.999
Prostatect.	FP	-0.072	-0.035	>0.999	0.339	>0.999
Prostatect.	NP	-0.071	-0.109	>0.999	0.253	>0.999
Hysterect.	FP	+0.021	+0.039	>0.999	0.339	>0.999
Hysterect.	NP	+0.018	+0.006	>0.999	0.508	>0.999

BH-adjusted p -values. All heterogeneity ($\hat{\alpha}_1$) and no-region tests non-significant.

A6a. Variable importance (no-region forests): Public vs PFP

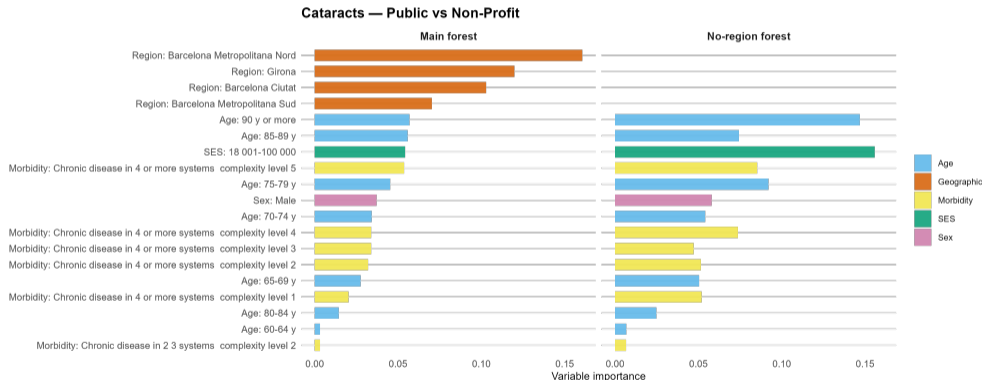
Variable Importance: Cataracts — Main vs. No-Region Forest

SES rises in relative importance when geographic dummies are excluded from the covariate set



Once region dummies are removed, age bands and SES rise to the top.

A6b. Variable importance (no-region forests): Public vs PNP



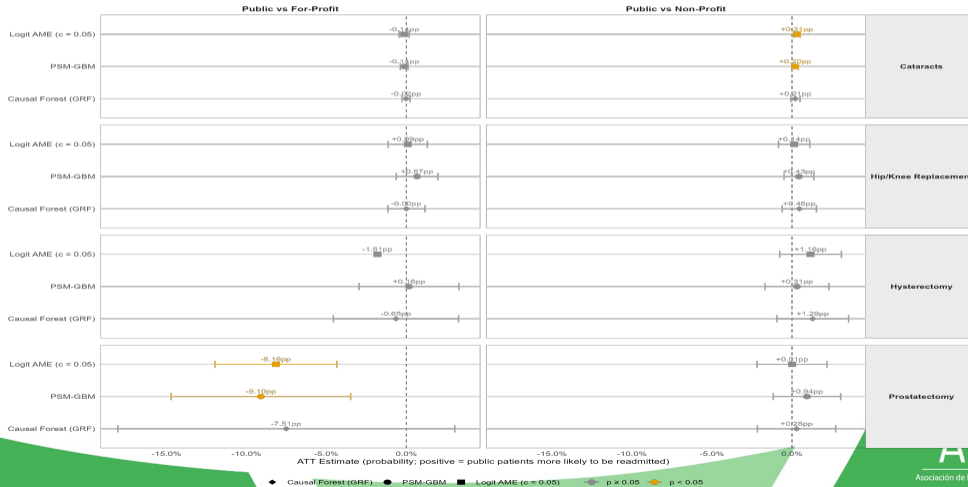
Horizontal bars: GRF variable importance scores. Color = variable type.
 Left panels: full covariate set including health region dummies.
 Right panels: covariate set without regio_sanitaria_hosp (region variables show importance = 0 by exclusion).

Pattern consistent across contrasts; the exploratory low-SES signal is not region-driven

A7. GBM vs. logistic matching: 30-day readmission

ATT Estimates for 30-Day Readmission — Method Comparison

Logit AME (c = 0.05) · PSM-GBM · Causal Forest (GRF) · probability scale



A8. Descriptive statistics by ownership & procedure

Procedure	Stat	ICS	Mun. Public	Other Public	Non-Profit	For-Profit
Cataracts	<i>N</i>	15,253	2,078	30,245	22,266	12,967
	LoS (d)	1.84	1.92	1.80	1.83	1.76
	Readm.	0.011	0.008	0.012	0.010	0.010
Hip & Knee	<i>N</i>	2,010	416	3,961	3,018	1,636
	LoS (d)	8.13	8.48	7.32	7.50	7.40
	Readm.	0.043	0.034	0.031	0.030	0.031
Hysterectomy	<i>N</i>	504	52	732	528	183
	LoS (d)	5.93	5.78	5.49	5.07	5.15
	Readm.	0.048	0 [†]	0.038	0.027	0.033
Prostatectomy	<i>N</i>	965	64	1,327	1,066	277
	LoS (d)	5.63	6.67	5.78	6.22	5.91
	Readm.	0.073	0.016	0.074	0.068	0.112

ICS = Institut Català de la Salut

(main public provider). Raw means; readmission = 30-day rate.

[†]Mun. public hysterectomy (*N* = 52): zero readmissions; excluded from readmission model.

A9. The AMG morbidity index (case-mix control)

- **Adjusted Morbidity Groups (AMG)**: a population morbidity index built by the Catalan health system (CatSalut) for risk stratification.
- Assigns each patient-year to **one of 31 mutually exclusive groups** from ICD-10 diagnoses recorded across *all* care settings during the year.
- Each group combines **disease burden** and **clinical complexity**, with up to five severity tiers; one ordinal-style measure of multimorbidity.
- Used routinely for **risk-adjusted capitation** in Catalonia; case-mix control for all three stages.

Monterde et al. (2016), *Aten. Primaria*; applied as in Maynou et al. (2023).

A10. Limitations

- **Statistical power beyond cataracts:** small matched samples for secondary procedures (hysterectomy–FP ≈ 143 treated) \Rightarrow hypothesis-generating only.
- **Referral endogeneity** (Decret 354/2002): private patients selected on capacity; unconfoundedness cannot be tested.
- **Inference:** < 30 clusters for several contrasts \Rightarrow cluster-robust CIs cautious.
- **Overlap:** severe in for-profit contrasts (up to 40% near boundary) \Rightarrow overlap-weighted estimand.
- **No quasi-experimental variation:** within-hospital capacity shocks or referral-rule discontinuities (Decret 354/2002) would identify ownership effects more cleanly.