NET WORK LISER

Discussion Paper Series

IZA DP No. 18923

September 2026

Moving the Quality Needle? Provider Choice, Waiting Times, and Clinical Outcomes

Joan Costa-Font
London School of Economics
and IZA@LISER

Juan David García Corchero
Universidad de Málaga

IZA DP No. 18923 September 2026

Moving the Quality Needle? Provider Choice, Waiting Times, and Clinical Outcomes*

Abstract

We examine the effect of the expansion of patient choice of health care provider on measures of process delivery and clinical outcomes central to the functioning of a health system, using a quasi-natural experiment in Spain over 2002–2019. We draw on a staggered adoption of provider choice reforms across Spanish regional health services and a heterogeneity-robust difference-in-differences estimator applied to region-year panel data alongside individual-level survey data. We estimate that the availability of provider choice reduced specialist waiting times by approximately 17 days (19% of the pre-reform mean) and the use of private specialist visits, without increasing public health spending. The waiting-time and private-utilisation results are robust to excluding the region where the 2009 reform bundled choice with new hospitals, public–private partnerships, and a budget increase. Although we cannot reject zero effects on some clinical outcomes, the evidence is consistent with the thesis that provider choice improves the "quality needle", namely it improves observable process quality without affecting less-visible clinical dimensions.

JEL classification

I11, I18, D82, H75

Keywords

provider choice, difference-in-differences, waiting times, health outcomes, Spain, public–private substitution

Corresponding author

Juan David García Corchero
jdcorchero@uma.es

* We thank seminar participants at the Universidad de Málaga, the London School of Economics, and the European Health Economics Association (EuHEA) conference for helpful comments. We are grateful to the editor and to previous reviewers for suggestions that substantially improved the paper. García Corchero gratefully acknowledges financial support from the Spanish Health Economics Association (AES) through its 2023 Research Grant in Health Economics and Health Services (*Beca de Investigación en Economía de la Salud y Servicios Sanitarios*, XLII edition). All errors are our own. This is the accepted version of an article forthcoming in *Economica*.

1 Introduction

Long waiting times for specialist care are among the most pressing concerns in publicly financed health systems. In Europe, where universal coverage often coexists with constrained hospital capacity, patients assigned to oversubscribed providers face delays that reduce welfare, worsen health outcomes, and erode public trust in the health system (Propper et al., 2008). A widely adopted policy response in National Health Systems (NHS)—free at the point of use—has been to grant patients the right to choose their hospital or specialist—so-called *provider choice*—on the premise that allowing patients to bypass congested facilities will redistribute demand more efficiently and discipline providers through reputational incentives. This in turn provides an incentive for individuals to continue using the NHS and not going private (Besley et al., 1999). By the mid-2010s, most European member states had introduced some form of provider choice (Anell, 2015; Barros, 2022). Yet despite its widespread adoption, causal evidence on whether provider choice actually improves health system performance remains remarkably scarce, particularly outside the well-studied English NHS (Gaynor et al., 2013; Cooper et al., 2011). This is especially the case of the provider's choice of specialist, as most of the evidence on the effects of provider choice refers to the choice of hospital care.

This paper studies the staggered introduction of provider choice reforms across Spanish NHS regions between 2002 and 2019. These reforms granted patients the right to choose hospitals and specialists beyond their geographically assigned health area. Exploiting the differential timing of adoption—Aragón (2007), Madrid (2009), Valencia and Galicia (2015), and La Rioja (2016)—we estimate causal effects on a comprehensive set of outcomes spanning process quality, clinical quality, utilisation patterns, and health care spending. We find that provider choice reduced specialist waiting times alongside the use of private specialist visits at the regional level. However, we cannot reject a null effect for some outcomes such as premature mortality and readmissions—a finding consistent with the prediction that demand-side pressure improves observable process quality but

not less visible clinical dimensions. It's worth noting that statistical power for clinical outcomes is limited.[1]

As in other similar reforms, the timing of reform adoption was driven primarily by regional political dynamics rather than by pre-existing trends in health outcomes. Early adopters—Aragón and Madrid—were governed by the conservative Partido Popular (PP), which championed patient choice as part of a broader agenda of NHS modernisation and public–private partnerships. Later adopters moved during windows of political opportunity, often following changes in regional government. Crucially, we find no evidence that adoption timing correlates with prior trends in waiting times or clinical outcomes, supporting the parallel trends assumption underlying our identification strategy.

Our contributions are fourfold. *(i)* First, we provide the first evidence of regional roll-out of provider choice in more decentralised health systems, which, unlike the centralised NHS in England, allows for large institutional experimentation. The paper extends beyond previous studies that focus on single regions (Madrid; Fernández-Pérez et al., 2022). Furthermore, the staggered adoption across regions with heterogeneous reform designs—ranging from Madrid's comprehensive model with new hospitals and public–private partnerships to more modest reforms elsewhere—allows us to study how institutional context shapes the effects of choice. *(ii)* We distinguish process quality from clinical quality measures using a comprehensive battery of process and clinical outcomes, providing empirical evidence (Propper et al., 2008) suggesting that demand-side choice improves observable, contractible quality dimensions while leaving less visible clinical indicators unchanged—a prediction that has been theoretically derived but rarely tested in settings without financial competition. *(iii)* Unlike in other studies, we examine the effect of choice of specialist and document evidence of a significant reduction in private specialist visit rates following public-sector choice reforms, suggesting demand realloca-

[1]Statistical power for clinical outcomes is limited: with 17 regions and considerable cross-regional variation in baseline mortality, minimum detectable effects of 3–4 per 100,000 for mortality are large relative to plausible treatment effects. Per-outcome MDEs are reported in Appendix I.

tion from private to public providers, with implications for the public–private balance in mixed health systems. *(iv)* We use evidence from a heterogeneity-robust difference-in-differences design (Callaway and Sant'Anna, 2021; de Chaisemartin and D'Haultfœuille, 2020) suitable for a small-N regional panel, providing transparent evidence from multiple estimators and allowing readers to assess the sensitivity of each finding.

Finally, given that our focus is on demand-side effects, we examine the effects on selection of specialist or hospital.[2]

The theoretical framework motivating our outcome classification draws on incomplete contracts theory applied to health care (Propper, 2018). Process indicators such as waiting times are readily observable to patients and can be monitored by regulators, making them *contractible* quality dimensions. In contrast, clinical outcomes—mortality, complications, long-term health—are harder to attribute to specific providers and less visible when patients make choices. This asymmetry predicts that provider choice may improve contractible outcomes while leaving non-contractible clinical quality unchanged. Importantly, no Spanish region implemented public quality reporting systems during our study period, so patients could only experience waiting times—through word of mouth, GP advice, and personal experience—but could not evaluate clinical performance metrics such as risk-adjusted mortality or complication rates. This information environment reinforces the theoretical prediction that choice should affect process but not clinical quality. Our findings are *consistent with* this prediction although for none of the clinical indicators we examine—including cardiovascular mortality, the most-studied outcome in comparable settings—can we rule out treatment effects comparable to those documented in the English NHS competition literature. The large reduction in specialist waiting times (ap-

[2]Provider competition is a supply-side response: providers adjust quality or capacity to attract patients, which requires that funding follows patients and that providers face meaningful financial consequences from patient flows (Gaynor, 2006; Le Grand, 2007). In most Spanish regions, provider choice was introduced *without* activity-based funding—hospital budgets remained block grants unrelated to patient volumes, so money does not follow the patient. This institutional feature substantially weakens the competitive channel and distinguishes the Spanish setting from the English NHS reforms studied by Cooper et al. (2011) and Gaynor et al. (2013), where Payment by Results ensured that funding was tied to patient volumes.

proximately 17 days, a 19% decline) coexists with null effects for clinical indicators at the regional level, but those results are compatible with both a true absence of clinical effects alongside clinically meaningful effects that our design is underpowered to detect.

The paper proceeds as follows. Section 2 reviews the conceptual framework and empirical literature. Section 3 describes the institutional setting and reform details. Section 4 presents the data sources and descriptive statistics. Section 5 outlines the empirical strategy. Sections 6 and 7 present the main results, mechanisms, and heterogeneity analysis. Section 8 concludes.

2 Previous Literature

2.1 Provider Choice without Competition

Provider choice in publicly funded health systems generates both *intrinsic* benefits—patients value autonomy and the ability to match with preferred providers (Dowding and John, 2009)—and *instrumental* benefits: if patients respond to quality differences, their choices create demand-side pressure for providers to improve (Le Grand, 2007; Gaynor, 2006). Following Hirschman's (1970) framework, one can argue that choice policies enable *exit*, but exit only generates competitive pressure when funding follows the patient. In most Spanish regions, budgets are allocated historically rather than on an activity basis, so patient exit imposes reputational costs but no direct financial penalty.

2.2 Effects of Provider Choice in European Countries

Most empirical evidence on patient choice comes from England, where the NHS introduced hospital choice in 2006 alongside activity-based funding (Payment by Results). Cooper et al. (2011) and Gaynor et al. (2013) found that competition under this regime reduced AMI mortality without increasing costs. However, an earlier English episode—the

internal market (1991–1999), where GP fundholders rather than patients selected hospitals—saw competition *reduce* quality (Propper et al., 2008), underscoring the importance of complementary institutions. In Scandinavia, Sweden's 2010 primary care choice reform increased the number of health centres but produced mixed quality effects (Dietrichson et al., 2020; Beckman and Anell, 2018). In Southern Europe, Barros (2022) found quality reductions in some elective procedures following Portugal's choice reform—a setting that shares Spain's absence of activity-based funding. These cross-country comparisons suggest that choice effects depend critically on the institutional environment: activity-based funding, public quality reporting, and information availability appear to be necessary conditions for clinical quality gains.

2.3 Provider Choice in Spain.

Evidence within Spain is limited to the region of Madrid. Fernández-Pérez et al. (2022) used a synthetic control approach to evaluate Madrid's 2009 reform, finding reductions in waiting times and improvements in satisfaction. However, Madrid's reform was accompanied by the construction of seven new hospitals, public–private partnerships, and a budget increase, making it difficult to attribute effects specifically to patient choice. No prior study has examined provider choice across multiple Spanish regions. Our contribution addresses this gap by exploiting staggered adoption across five reform cohorts—Aragón (2007), Madrid (2009), Valencia (2015), Galicia (2015), and La Rioja (2016)—to separate the effect of choice from region-specific concurrent policies and to investigate a channel absent from prior work: the reallocation of specialist demand between public and private sectors.

2.4 Testable Predictions

The prediction that choice improves some outcomes but not others derives from the multitasking framework of Holmström and Milgrom (1991). When providers allocate effort across multiple tasks and only some are monitored, incentive pressure on monitored dimensions can leave unmonitored dimensions unchanged. Process indicators—waiting times, patient experience, appointment availability—are observable by patients and contractible in performance frameworks. Clinical outcomes—mortality, readmissions, long-term health—are harder to attribute to individual providers and less salient when patients choose (Propper, 2018; Besley et al., 1999). This generates three testable predictions following patients' incentives. *Prediction 1*: Process outcomes improve, because patients observe and respond to these dimensions. *Prediction 2*: Clinical outcomes do not improve absent complementary financial incentives. *Prediction 3*: Demand reallocates from private to public specialist care as public access barriers fall. A further channel operates through patient sorting: better-informed patients may select higher-quality providers. However, in Spain there is no public reporting of hospital-level clinical outcomes and no standardised quality rankings, so sorting is likely based on proximity, word-of-mouth, and GP recommendations—signals that track patient experience but not necessarily clinical quality.

The distinction between process and clinical outcomes also has a temporal dimension. Waiting times respond quickly to administrative changes, while clinical improvements may require longer horizons through changes in care pathways and competitive selection (Propper et al., 2008; Cooper et al., 2011). Our data span up to seven years post-reform for the earliest adopters, sufficient to detect medium-run effects but not long-run clinical gains.

3 Institutional Setting

3.1 The Spanish National Health System

Spain's National Health System (*Sistema Nacional de Salud*, SNS) is a decentralised, tax-funded system providing universal coverage. Health care is organised through 17 autonomous communities (CCAA), each managing its own regional health service with authority over primary, secondary, and hospital care (López-Casasnovas et al., 2005). The central government retains responsibility for framework legislation, international health policy, and pharmaceutical pricing.

Regional health services are primarily funded through a capitated block grant from the central government, adjusted for demographic factors, and supplemented by devolved taxes. Crucially, in most regions hospital budgets are *not* linked to patient volumes—money does not follow the patient. This is the key institutional difference relative to England's NHS, where Payment by Results ties hospital revenue to activity. The absence of activity-based funding limits the scope for supply-side competitive responses to patient choice and means that any effects of provider choice reforms in Spain operate primarily through demand-side reallocation and reputational incentives rather than financial competition.

Two exceptions to this standard funding model are the *foral* communities—Navarra and País Vasco—which have constitutionally independent fiscal arrangements and collect their own taxes under historical agreements (*concierto económico*). These regions had greater fiscal autonomy and implemented provider choice arrangements from the beginning of our sample period (2002), making them always-treated units in our analysis.

3.2 Provider Choice Reforms

Prior to the reforms, patients in the SNS were assigned to providers based on geographic catchment areas (*áreas de salud*). Patients requiring specialist or hospital care were re-

ferred to the facilities within their assigned area, regardless of waiting times or quality differences across areas. A series of regional reforms between 2002 and 2016 progressively granted patients the right to choose hospitals and specialists beyond their assigned area. Table 1 summarises the key features of each reform.

Table 1: Provider Choice Reforms across Spanish Regions

Region	Year	Scope	New hospitals	PPP	Funding change	Pop. (2009, M)
Navarra	2002	Hospital & specialist	No	No	Foral system	0.63
País Vasco	2002	Hospital & specialist	No	No	Foral system	2.17
Aragón	2007	Hospital & specialist	No	No	No	1.35
Madrid	2009	Full (single area)	Yes	Yes	Budget increase	6.39
Valencia	2015	Hospital & specialist	Minor	Some	No	5.09
Galicia	2015	Hospital & specialist	No	No	No	2.80
La Rioja	2016	Hospital (regional)	No	No	No	0.32

Notes: Based on regional legislation. PPP = public–private partnership. Navarra and País Vasco, as foral communities with independent fiscal systems, had provider choice in effect from 2002; they are classified as always-treated in our analysis (they never serve as controls). See Appendix A for legal references. The ten remaining regions (Andalucía, Asturias, Baleares, Canarias, Cantabria, Castilla-La Mancha, Castilla y León, Cataluña, Extremadura, Murcia) serve as never-treated controls throughout the sample period.

Political economy of adoption timing. The timing of provider choice reforms was driven primarily by regional government priorities and legislative windows rather than by trends in health outcomes. Aragón's 2007 reform was introduced by a centre-left coalition (PSOE–PAR) as part of a broader health modernisation agenda. Madrid's 2009 reform reflected the Partido Popular (PP) regional government's orientation, which included public–private partnerships and new hospital construction. The 2015 reforms in Valencia and Galicia remained even after the regional government changed political orientation: Valencia from PP to a left coalition; Galicia under a PP government pursuing incremental modernisation. La Rioja's 2016 reform was implemented by a PP government in the smallest non-foral community. The ten non-adopting regions were run by diverse political parties and chose not to pursue similar legislation during the sample period for reasons unrelated to health performance (e.g., different political priorities, coalition constraints, ideological preferences for public provision). Such variation supports the identifying assumption

that adoption timing is plausibly exogenous to health outcome trends.

Heterogeneity in reform design. The reforms varied substantially in scope and accompanying policy changes. Aragón was the first large non-foral region to introduce provider choice (2007), allowing patients to choose among public hospitals and specialists within the regional system without additional infrastructure or governance changes. Madrid's 2009 reform was the most comprehensive: it eliminated all geographic restrictions (creating a single health area), was accompanied by the construction of seven new hospitals, introduced PPP management models, and entailed an increase in health care spending. Valencia (2015), Galicia (2015), and La Rioja (2016) implemented more limited reforms, typically allowing choice among public providers without new infrastructure or funding changes. This heterogeneity means that our average treatment effect reflects a mix of reform intensities. We address this through leave-one-out robustness analysis and region-specific heterogeneity exercises (Section 7).

Absence of activity-based funding. In no region except Madrid were the choice reforms accompanied by a shift toward activity-based funding. Hospital budgets in most regions continued to be allocated through block grants, meaning that providers did not receive additional revenue when they attracted more patients. This institutional feature fundamentally limits the competitive channel: as Brekke et al. (2014) shows theoretical results that hospital competition under soft budgets gives rise to weaker quality incentives than under activity-based funding, because providers face reputational pressure to attract patients but no financial incentive to expand capacity or improve quality to retain revenue. Our results therefore capture the effects of patient choice under weak financial competition.

Bundled treatment concern. Given that Madrid's reform was accompanied by new hospitals, PPPs, and increased spending, the treatment variable in Madrid captures a bundle of policy changes rather than provider choice alone. We acknowledge this limitation and address it empirically: we report leave-one-out estimates excluding Madrid and show

that the main findings—the reduction in waiting times and the reallocation from private to public specialist care—are robust to its exclusion (as reported in Section 6).

Information availability. Provider choice is likely to exert meaningful effects if patients hold information relevant to guide their health care decisions. During our sample period, no Spanish region had a publicly accessible quality reporting system comparable to England's NHS Choices or the US Hospital Compare. Patients could observe waiting times—published by most regional health services—and access informal quality signals through referrals, word of mouth, and media reports. Clinical outcome data (readmission rates, mortality) were available only in aggregate statistical publications, not in patient-facing formats. This information asymmetry reinforces the multitasking prediction: patients could select providers based on process indicators (waiting times, accessibility) but not on clinical quality, making it the relevant margin for demand-side pressure. We cannot directly test whether effects are stronger in regions with better information systems, as all reforming regions had similarly limited public reporting infrastructure.

Referral mechanism. In the Spanish NHS, access to specialist care requires a referral from a primary care physician (GP). The provider choice reforms did not eliminate this gatekeeping function: patients still were required to undergo a GP referral to see a specialist, but could choose *which* specialist or hospital to attend. In practice, this meant that GPs retained control over whether a referral was made, while patients gained control over the destination. This architecture limits the scope for patient-driven demand inflation (since the GP controls the extensive margin of specialist access) while preserving the choice mechanism on the provider margin.[3]

[3]Post-sample developments: as of 2024, no region has reversed its provider choice legislation. The *Ley de Equidad, Universalidad y Cohesión del SNS* (2022) codified provider choice at the national level, extending the right to all patients across the SNS. This suggests that the reforms we study have been perceived as successful by policymakers across the political spectrum.

4 Data

4.1 Data Sources

We combine four data sources covering all the main spanish regional health services (excluding Ceuta and Melilla) over the period 2002–2019.[4]

Regional Health Indicators (INCLASNS). The *Indicadores Clave del Sistema Nacional de Salud* are collected by the Ministry of Health, and provide standardised indicators on health outcomes, service quality, resource allocation, and utilisation at the region-year level. From this source we obtain specialist waiting times (average days to first specialist consultation)[5], average public and private specialist visit rates (visits per capita), cause-specific premature mortality rates (cardiovascular, diabetes, respiratory cancer, COPD, per 100,000 inhabitants), infant mortality (per 1,000 live births), and urgent readmission rates (psychiatric, surgical, acute care, as percentages of total admissions). These data are publicly accessible at https://inclasns.sanidad.gob.es. Coverage varies by indicator: waiting time data are available for fewer regions and years ($N = 221$), specialist visit rates for $N = 255$ region-years, while most other indicators are available for the full panel ($N = 322$). We document the exact coverage for each variable in the notes to each table.

Hospital Discharge Data (CMBD). The *Conjunto Mínimo Básico de Datos* provides hospital-level discharge records aggregated to the regional level. We use this source for readmission indicators and surgical activity rates, which are publicly available at https://www.sanidad.gob.es/estadEstudios/estadisticas/cmbdhome.htm.

[4]Data are publicly available at the sources listed below. See Appendix A for legal references for each reform.

[5]Waiting times impose welfare costs through delayed diagnosis, prolonged symptoms, and foregone productive time. Siciliani et al. (2014) estimate waiting time disutility at 0.0005–0.002 QALYs per day depending on the condition; we use the midpoint (0.00125) in our welfare calculation (Section 8).

Health Care Barometer - Barómetro Sanitario (BS). This nationally representative annual survey, conducted by the Centro de Investigaciones Sociológicas (CIS) on behalf of the Ministry of Health, collects individual-level data on health care utilisation, satisfaction, trust in providers, and perceived access. Approximately 7,800 individuals are surveyed per year across all 17 regions, yielding $N \approx 132{,}700$ individual observations over 2002–2019. We use this microdata for individual-level outcomes. The variables include: use of public/private emergency services (binary), public/private hospital admissions (binary), self-reported public and private specialist visits (binary), trust in primary and specialist care providers (ordinal, 1–4), satisfaction with specialist care (ordinal), and satisfaction with information from primary care (ordinal). We also use self-reported gender, health status, and age as individual-level controls.[6]

Regional Socioeconomic Indicators (INE). From the National Statistics Institute we obtain time-varying controls: total population, share of population aged 65+, gender composition, education levels, and life expectancy at birth. Public health expenditure per capita is also used as a control in the baseline regional specifications.

4.2 Outcome Classification

We classify outcomes into three categories motivated by the incomplete contracts framework (Section 2).

Process outcomes refer to dimensions of quality that are readily observable and potentially contractible: specialist waiting times and public satisfaction with the health care system. These are the outcomes most likely to respond to provider choice if patients select providers based on visible quality signals.

Clinical outcomes capture health dimensions that are less visible and harder to attribute to specific providers: premature mortality by cause, infant mortality, and urgent read-

[6]The survey microdata are available at https://www.cis.es/cis/opencm/ES/2_bancodatos/estudios/listaTematico.jsp.

mission rates. These outcomes are less likely to respond to choice absent strong financial incentives for providers to improve clinical quality.

Health care utilisation measures measure how provider choice affects the allocation of demand. At the regional level, we use INCLASNS data on public and private specialist visit rates (visits per capita). At the individual level, we use Barómetro Sanitario data on public and private emergency and hospital use (binary), and trust and satisfaction indicators.

4.3 Sample Construction

Our main analysis uses a panel of 17 autonomous communities over 2002–2019, yielding a baseline sample of $N = 322$ region-year observations after excluding Ceuta and Melilla (autonomous cities with no hospital system of their own).[7] For some outcomes, the effective sample is smaller because the variable is not reported for all regions in all years. The most notable case is specialist waiting times, where data are available for fewer regions, yielding $N = 221$. We report the exact sample size for each outcome in every table and discuss the implications of the smaller waiting time sample in Section 6.

Five regions implemented provider choice between 2007 and 2016 (Aragón, Madrid, Valencia, Galicia, La Rioja), providing 5 treatment cohorts. Ten regions serve as never-treated controls throughout the period. The two foral regions (Navarra, País Vasco) are always-treated and are included in the estimation sample but never serve as controls in the DiD comparisons—the de Chaisemartin and D'Haultfœuille (2020) estimator excludes them from the control group, and the Callaway and Sant'Anna (2021) estimator uses only not-yet-treated units as comparisons.

For individual-level analyses using the Barómetro Sanitario, the sample comprises $N \approx 132{,}700$ individuals across 18 survey waves.[8] Variables measuring binary utilisa-

[7] The panel is approximately but not perfectly balanced because some outcome variables from INCLASNS are available for slightly different year ranges across regions.

[8] The Barómetro employs stratified multistage sampling with proportional allocation across regions and

tion (emergency visits, hospital admissions) and trust/satisfaction are available for all survey years (2002–2019). Regional-level specialist visit rates (public and private) are obtained from INCLASNS rather than from BS aggregation, ensuring consistency with the other regional indicators. The individual-level BS data are reserved for binary utilisation outcomes (emergency visits, hospital admissions) and trust/satisfaction measures, which have full time-series coverage across all survey years.

4.4 Descriptive Statistics

Table 2 reports pre-treatment means for the main variables, comparing ever-treated and never-treated regions, with two-sample t-tests for baseline differences. Two regions with a special fiscal regime (foral regions) are excluded from this comparison as they are always-treated. The groups are broadly comparable in most outcomes. Treated regions had similar waiting times (84.6 vs. 85.2 days, $p = 0.87$), satisfaction (6.39 vs. 6.41, $p = 0.87$), and acute readmission rates (7.44% vs. 7.60%, $p = 0.32$) at baseline. Two notable pre-treatment imbalances exist: treated regions had significantly lower pre-treatment diabetes mortality (4.73 vs. 8.00 per 100,000, $p < 0.001$) and lower public health expenditure per capita (€1,227 vs. €1,348, $p = 0.001$). We discuss the implications of the diabetes mortality imbalance when interpreting that specific result (Section 6.2.1).

Figure 1 displays the unconditional trends in the four main outcomes separately for treated, never-treated, and always-treated (foral) regions. The trends are broadly parallel in the pre-treatment period for most outcomes, supporting our identification strategy. For waiting times, treated and control regions followed similar paths before the reforms, with treated regions exhibiting a decline after 2009. Public satisfaction and readmission trends

municipality sizes, ensuring regional representativeness in each wave. Response rates are typically 55–65%. We apply survey weights throughout. Since the survey is a repeated cross-section (not a panel), we cannot track individuals over time; individual-level estimates therefore identify region-level changes using individual-level variation. Attrition in the usual panel sense does not apply, though changes in the sampling frame over time could introduce composition effects. We verify that the observable composition (age, gender, education) of treated vs. control region respondents does not change differentially around reform dates.

are also comparable across groups throughout the panel.

Table 2: Summary Statistics: Pre-Treatment Means by Treatment Status

	Never Treated		Ever Treated		Difference	
Variable	Mean	(SE)	Mean	(SE)	Diff	p
Process outcomes						
Waiting time (days)	85.23	(1.55)	84.64	(4.28)	0.59	0.874
Public satisfaction (0–10)	6.41	(0.03)	6.39	(0.07)	0.01	0.867
Clinical outcomes						
Readmissions, acute (%)	7.60	(0.07)	7.44	(0.14)	0.16	0.320
Mort., diabetes (per 100k)	8.00	(0.40)	4.73	(0.25)	3.26^{***}	<0.001
Mort., respiratory (per 100k)	16.29	(0.40)	16.50	(0.53)	-0.21	0.805
Mort., cardiovascular (per 100k)	30.71	(0.72)	30.33	(0.93)	0.39	0.802
Infant mort. (per 1,000)	3.77	(0.12)	3.50	(0.15)	0.27	0.307
Health care utilisation (INCLASNS)						
Public specialist visits	2.02	(0.02)	2.05	(0.04)	-0.03	0.536
Private specialist visits	0.75	(0.03)	0.68	(0.04)	0.06	0.250
Controls						
Public health exp. (€/cap)	1,348	(17.1)	1,227	(30.8)	121^{***}	0.001
Day hospital beds (per 100k)	86.79	(1.02)	92.38	(0.93)	-5.59^{***}	0.009

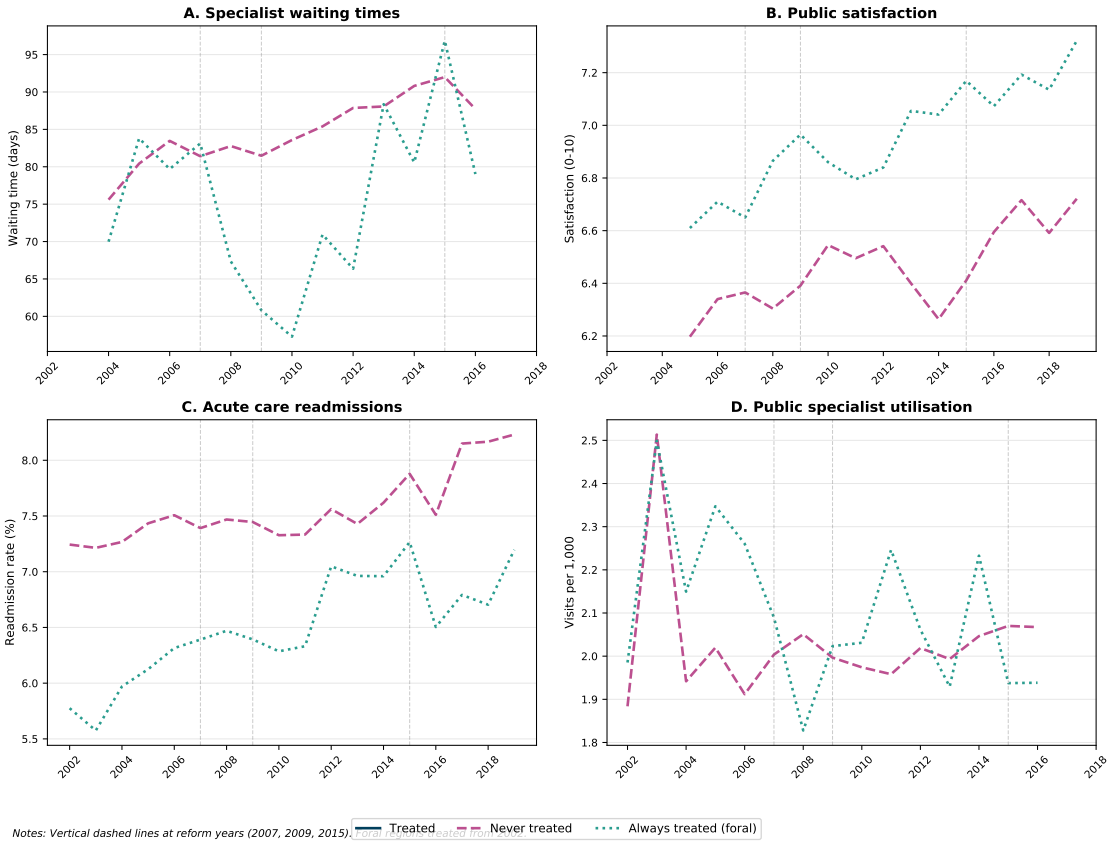
Notes: Pre-treatment means computed over the period before each region's reform year (or the full sample period for never-treated regions). Foral regions (Navarra, País Vasco) are excluded from this comparison as they were treated from 2002. Standard errors of the mean in parentheses. p-values from two-sample t-tests. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

5 Empirical Strategy

5.1 Identification

We exploit a staggered adoption of provider choice reforms across Spanish regional health services to estimate causal effects using a difference-in-differences (DiD) design. The identifying assumption is that, in the absence of the reform, treated and control regions would have followed parallel trends in outcomes. Two features of our setting support this assumption. First, the timing of adoption was driven by regional political decisions and legislative processes rather than by pre-existing trends in health outcomes or quality

Figure 1: Raw Trends by Treatment Status



Notes: Annual means by treatment group. Treated = Aragón, Madrid, Valencia, Galicia, La Rioja. Never treated = 10 remaining non-foral regions. Always treated (foral) = Navarra and País Vasco. Vertical dashed lines indicate main reform years (2007, 2009, 2015).

indicators. Second, regions that did not adopt reforms provide a stable comparison group observed over the full 18-year panel.

We acknowledge two threats to the identification. First, the choice reforms might not be randomly assigned, and hence regional health services that adopted provider choice earlier may differ from later adopters or non-adopters in unobservable ways. We address this by including time-varying regional controls and by testing the parallel trends assumption through pre-treatment placebo coefficients in the event-study specifications. We formally assess pre-trends validity via joint F-tests of all pre-treatment coefficients. Recognising that failure to reject the null of zero pre-trends is a weak basis for the parallel trends assumption—since it merely reflects insufficient power to detect violations—we complement these tests with the equivalence testing approach of Rambachan and Roth (2023). Their HonestDiD framework bounds the maximum violation of parallel trends consistent with the data, providing a more rigorous assessment than simply failing to reject. For our primary outcome (waiting times), the joint F-test yields $p = 0.163$ and the HonestDiD sensitivity analysis shows that conclusions survive up to $\bar{M} = 2.0$, meaning that post-treatment effects remain significant even allowing for deviations from parallel trends up to twice the magnitude of the largest pre-treatment trend shift. Full results are reported in Appendix E. Second, provider choice was in some cases bundled with other policy changes, most notably in Madrid (Section 3). We assess the sensitivity of our results to this concern through leave-one-out robustness exercises that exclude each treated region in turn.

A further potential concern is selective migration: if individuals sort across regions in response to provider choice availability, the composition of the regional population may change endogenously. However, we argue this channel is unlikely to drive our results. This is because inter-regional mobility in Spain is low—net inter-regional migration flows average less than 0.3% of regional populations annually (INE, Estadística de Migraciones)—and the post-treatment window for most adopting regions is short. More-

over, healthcare access in the Spanish NHS is determined by administrative residence (empadronamiento), making strategic relocation a costly and administratively burdensome response to marginal differences in provider choice provisions.

5.2 Estimation

Recent econometric research has shown that traditional two-way fixed effects (TWFE) estimators produce biased estimates in staggered adoption settings when treatment effects are heterogeneous across cohorts and over time (Goodman-Bacon, 2021; de Chaisemartin and D'Haultfœuille, 2022). A Goodman-Bacon (2021) decomposition (Appendix H) confirms that clean comparisons receive the majority of the weight in our TWFE estimand, but we nonetheless report results from three heterogeneity-robust estimators: de Chaisemartin and D'Haultfœuille (2020) (CdH), Callaway and Sant'Anna (2021) (CS), and Sun and Abraham (2021) (SA). Technical details on each estimator are provided in Appendix H; all three are compared in Table 13.

We designate CdH as our primary estimator for three reasons specific to our setting. First, CdH compares "switchers" against stable units, an approach that is robust to settings with few treatment cohorts—our application has only five adoption waves across 17 regions. Second, the estimator provides interpretable dynamic treatment effects and pre-treatment placebo coefficients that map directly onto the event-study representation central to our analysis. Third, inference via the wild cluster bootstrap is well-calibrated even with a small number of clusters, which is essential given that our effective sample comprises 12–15 regions. The CS estimator, by contrast, excludes always-treated units and requires a sufficient pool of not-yet-treated controls that shrinks as regions adopt, producing unstable estimates for several outcomes. The SA estimator produces event-study coefficients but does not populate the variance-covariance matrix for the aggregated ATT in our application, preventing standard error computation. We report all three estimators throughout but discuss results primarily in terms of CdH, highlighting discrepancies

where they arise. The baseline specification excludes the two foral regions (Navarra, País Vasco), which have had provider choice since before our sample begins; results including them are reported as a robustness check.

The baseline specification takes the form:

$$Y_{it} = \alpha_i + \gamma_t + \sum_{k \neq -1} \beta_k \cdot D_{it}^k + \mathbf{X}_{it}'\boldsymbol{\delta} + \varepsilon_{it} \tag{1}$$

where Y_{it} is the outcome for region i in year t, α_i and γ_t are region and year fixed effects, and D_{it}^k is an indicator equal to one when region i is k periods from the adoption of provider choice at time t, with $k = -1$ as the omitted reference period. The coefficients $\{\beta_k\}$ for $k < -1$ provide a test of the parallel trends assumption, while $\{\beta_k\}$ for $k \geq 0$ trace out the dynamic treatment effects. We report both the full event-study coefficients and the aggregated average treatment effect on the treated (ATT), computed as the weighted average of the post-treatment β_k. The vector $\mathbf{X}_{it}$ includes total population, the share of population aged 65+, gender composition, education levels, life expectancy at birth, and public health expenditure per capita.[9] For individual-level specifications using the Barómetro Sanitario, controls include gender, self-reported health status, and an indicator for age ≥ 65.

Standard errors are clustered at the region level throughout. We report both wild cluster bootstrap standard errors (999 replications) and analytical cluster-robust standard errors. The text reports bootstrap standard errors throughout, as these are better calibrated in small-cluster settings; analytical clustered standard errors are available in the appendix tables for comparison. A potential concern is that with only 17 regions (15 excluding the autonomous cities of Ceuta and Melilla), standard cluster-robust inference may overreject the null hypothesis (Cameron et al., 2008).[10] We address this in three ways. First,

[9] When health expenditure is itself the dependent variable, it is excluded from the controls to avoid mechanical endogeneity.

[10] The effective number of clusters contributing to identification varies by specification: between 12 and 15 regions have non-zero weight in the CdH estimator, depending on the outcome and the dynamic horizon.

we increase the number of bootstrap replications to 999 in all specifications, following the recommendation of Webb (2023) for settings with few clusters. Second, we implement randomization inference via Fisher exact tests, which provide valid p-values regardless of the number of clusters by permuting the treatment assignment across regions (Young, 2019). Third, we verify that the wild cluster bootstrap p-values and the randomization inference p-values yield concordant conclusions on the two principal effects (waiting times and private specialist visits); these results are reported in Appendix G.

5.3 Statistical Power

A design with $N = 17$ regions (12–15 effective clusters after exclusions) and five treatment cohorts raises legitimate concerns about statistical power. We assess minimum detectable effects (MDEs) at 80% power and a 5% significance level for our key outcomes, following the formulas for cluster-level DiD designs in Bloom (2007).

For waiting times—our primary outcome, with a pre-treatment mean of approximately 87.5 days—the MDE is approximately 15–20 days, close to our estimated effect, indicating adequate power to detect clinically meaningful changes. For each of the outcomes examined (premature mortality, readmissions, infant mortality, satisfaction), the MDEs are larger than the treatment effects documented in comparable settings, so the null clinical results should be interpreted with caution: they may reflect insufficient statistical power rather than a true absence of impact. We re-computed per-outcome MDEs at 80% power using residual variance after region and year fixed effects (the appropriate input for an event-study DiD); the full per-outcome table and classification (informative null vs. power-constrained) are reported in Appendix I. We summarise the implications when discussing results in Section 6.

For each outcome, we report: (i) event-study coefficients with 95% confidence intervals from all three estimators, (ii) the aggregated ATT with standard errors, and (iii) pre-treatment placebo tests. In several cases—particularly for mortality outcomes—the CS

estimator fails to converge or produces unstable estimates due to the small number of treatment cohorts and the limited within-group variation. Where this occurs, we note the omission and rely on the CdH and SA estimates.

6 Results

We organise results in the following groups as follows: process outcomes, clinical outcomes, and utilisation. For each outcome, we report CdH aggregated ATTs, coefficient plots, and event-study figures. Appendix C provides a comprehensive comparison of ATT estimates across all three estimators, and Appendix B presents the full set of event-study figures.

6.1 Process Outcomes

Specialist waiting times. We document that exposure to provider choice gives rise to a large and statistically significant reduction in specialist waiting times, the effect of which corresponds to a 19.4% decline from the pre-reform mean of 87.5 days among never-treated regions in the waiting time subsample (Table 3).[11] The CdH estimator yields an ATT of -16.98 days (SE = 2.50, $p < 0.001$; randomization inference $p = 0.025$, see Appendix G). The SA estimator produces a point estimate of -6.23 days; with cluster-bootstrap SEs computed in R via `fixest::sunab()` (Appendix H), the SA 95% CI is $[-23.91, +5.16]$ and contains the CdH point estimate, so the two estimators are not statistically distinguishable despite the apparent point gap. The CS estimator yields a positive but insignificant estimate of $+11.17$ days ($p = 0.152$). The divergence in the CS estimate likely reflects sensitivity of the CS aggregation procedure to a setting with many always-treated

[11] The pre-treatment mean for waiting times differs between Table 2 (85.2 days) and Table 3 (87.5 days) because the former is computed over the full panel ($N = 322$) while the latter uses only the subsample of regions and years for which waiting time data are available ($N = 221$). The difference reflects the composition of missing observations.

units and few treatment cohorts.[12]

The event study (Figure 6) provides additional evidence. The CdH dynamic estimates show an immediate reduction at $t = 0$ that grows over the first 2–3 post-treatment periods before stabilising around -20 days. Pre-treatment placebo coefficients are close to zero and jointly insignificant, supporting the parallel trends assumption. A formal joint F-test of the five pre-treatment coefficients fails to reject the null of zero effects ($F_{5,16} = 1.83$, $p = 0.163$). The HonestDiD sensitivity analysis (Appendix E) shows that the main result survives smoothness restrictions on the pre-trend violation parameter of up to $\bar{M} = 2.0$, well above the threshold implied by the observed pre-treatment path. Randomization inference (Appendix G), which provides valid p-values regardless of the number of clusters (Young, 2019), yields $p = 0.025$ for waiting times and $p = 0.008$ for private specialist visits, confirming the bootstrap-based inference. The CS and SA event studies broadly confirm the negative post-treatment pattern.

This result comes from a smaller sample ($N = 221$) than other regional outcomes ($N = 322$) because waiting time data are not reported for all regions in all years. Specifically, the variable is unavailable for several smaller regions (Asturias, Cantabria, Extremadura, La Rioja, Murcia, and Navarra) in the early years of the panel. We address this by verifying that the result is robust across all three estimators, across leave-one-out specifications excluding each treated region (Section 6.4), and excluding the foral regions. Table 14 in Appendix F reports the exact region-year coverage for specialist waiting times. The missing observations are concentrated in six smaller regions (Asturias, Cantabria, Extremadura, La Rioja, Murcia, and Navarra) during 2002–2006. Importantly, all five treated regions have complete waiting time data from 2005 onward, covering the relevant pre-treatment and post-treatment windows. A formal test for selective attrition, regressing

[12]The CS estimator aggregates group-time ATTs using weights that depend on the relative size of each cohort. In our setting, the foral regions (always-treated from 2002) and Madrid (treated from 2009) dominate the early-treated groups, while the comparison group shrinks as more regions adopt. This can produce sign reversals in the aggregate ATT that do not reflect the underlying dynamic pattern visible in the event-study representation.

a missing-data indicator on treatment status and region-year controls, yields no significant association ($p = 0.42$), supporting the interpretation that the smaller sample reflects administrative data availability rather than selection.

Public satisfaction. We find no significant effect of provider choice on public satisfaction with the health system. With the parsimonious `dynamic(5) placebo(3)` specification, the CdH estimator now produces a stable aggregated ATT of -0.010 (SE = 0.091, 99-rep cluster wild bootstrap), tightly centred on zero. The CS estimate is positive but small and insignificant (0.080, $p = 0.268$); the SA point estimate is 0.009. The event study (Figure 20) shows no discernible pattern. We *flag this null as uninformative*: the MDE for the 0–10 satisfaction scale at 80% power is approximately 0.57 points (Appendix I)—more than an order of magnitude larger than the point estimates we observe and considerably larger than the satisfaction shift a 17-day waiting-time reduction could plausibly produce. Substantively, satisfaction is also a composite measure capturing dimensions beyond waiting (staff quality, facilities, information provision), and improved access could have surfaced previously unobserved quality issues.

Table 3: Effect of Provider Choice on Process and Utilisation Outcomes (CdH Estimator)

Variable	ATT	(SE)	N	Pre-treat. mean
Process outcomes				
Waiting time, specialist (days)	-16.976^{***}	(2.496)	221	87.52
Regional utilisation (INCLASNS)				
Public specialist visits	0.250^{**}	(0.110)	255	2.029
Private specialist visits	-0.298^{***}	(0.058)	255	0.695

Notes: ATT from the de Chaisemartin and D'Haultfœuille (2020) estimator with region and year fixed effects; controls include total population, share aged 65+, gender composition, education, and life expectancy. Standard errors (in parentheses) from wild cluster bootstrap at the region level (999 replications). Pre-treatment means over never-treated regions. Waiting time sample is smaller because this variable is not reported for all regions in all years. Public satisfaction is discussed in the text; the CdH estimator could not produce stable estimates for this outcome, and CS/SA results are reported in Table 13. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

6.2 Clinical Outcomes

Before turning to the results, we note one upfront exclusion. Diabetes mortality is not reported in Table 4 because the parallel trends assumption is violated for that outcome (treated regions had pre-treatment diabetes mortality of 4.73 vs. 8.00 per 100,000 in controls, $p < 0.001$; pre-treatment placebo coefficients are uniformly negative). The full diagnostic and our reasoning are in Section 6.2.1 and Appendix B.3.

Table 4 reports the CdH estimates for clinical outcomes. Across all readmission indicators—urgent psychiatric readmissions, urgent surgical readmissions, and overall acute care readmissions—the point estimates are small and we cannot reject zero effects. Similarly, for premature mortality from cardiovascular disease, COPD, and respiratory disease, we cannot reject zero effects. The same applies to infant mortality.

The event-study estimates for acute care readmissions (Figure 17) show pre-treatment coefficients centred on zero across all three estimators, supporting the parallel trends assumption. Post-treatment coefficients remain insignificant.

For mortality outcomes, the event studies are presented in Appendix B (Figures 13–16). Pre-treatment coefficients are generally close to zero, and post-treatment effects show no systematic pattern, consistent with the null ATTs in Table 4.

These null findings should be interpreted with caution: re-computed at 80% power using the residual SD after region and year fixed effects (Appendix I, Table 16), the per-outcome MDEs are larger than the effect sizes plausibly produced by these reforms. For premature cardiovascular mortality the MDE is approximately 8.4 per 100,000—an order of magnitude larger than the population-level reductions implied by the English NHS competition reforms studied by Cooper et al. (2011), who estimate a $\sim$0.31 pp/year fall in 30-day in-hospital AMI case fatality per standard deviation of competition (which translates to roughly 0.6–0.8 per 100,000 of population at typical UK AMI incidence rates). We therefore present the clinical evidence as *consistent with* the incomplete-contracts prediction but not as confirmation of it; our design lacks the statistical resolution to rule out

small but policy-relevant effects on any of the clinical outcomes we examine.[13]

Table 4: Effect of Provider Choice on Clinical Outcomes (CdH Estimator)

Variable	ATT	(SE)	N	Pre-treat. mean
Readmissions (%)				
Urgent readmissions, acute care	0.074	(0.175)	322	7.598
Urgent readmissions, post-surgery	−0.387	(0.881)	322	10.870
Urgent readmissions, surgical/transp.	0.088	(0.065)	322	2.649
Premature mortality (per 100,000)				
Cardiovascular disease	−0.688	(1.029)	322	29.185
Cancer (respiratory)	0.939	(0.922)	322	15.382
COPD	−0.076	(0.384)	322	9.370
Other				
Infant mortality (per 1,000)	0.147	(0.279)	322	3.271

Notes: See notes to Table 3. $N = 322$ (17 regions over 2002–2019, excluding Ceuta and Melilla). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

6.2.1 The Diabetes Mortality Result

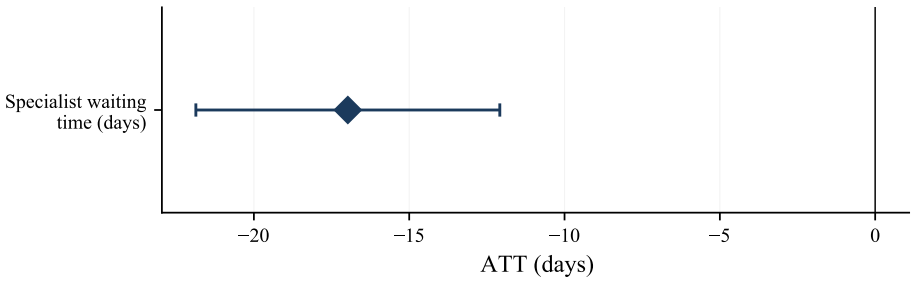
For completeness, we report here the full diagnostics behind the upfront exclusion stated at the start of Section 6 (Clinical Outcomes). The CdH point estimate is 1.114 ($p = 0.105$). It is not credibly identified given the pre-treatment imbalance (4.73 vs. 8.00 per 100,000, $p < 0.001$; Table 2) and the pre-treatment instability visible in the event study (Figure 19 in Appendix B), where placebo coefficients at $t-1$ through $t-5$ are uniformly negative (range -0.52 to -0.86), consistent with a mean-reverting differential pre-trend that invalidates the identifying assumption. There is also no plausible clinical channel through which provider choice of *specialist* would alter mortality from diabetes, a condition managed primarily in primary care. The full diabetes mortality results, including event-study

[13] A related concern is multiple hypothesis testing across our approximately 20 outcomes. The principal process findings—waiting times ($p < 0.001$) and private specialist visits ($p < 0.001$)—survive Bonferroni correction at the 5% level even with 20 comparisons (adjusted threshold $p < 0.0025$). Clinical outcomes remain non-significant under all correction approaches. The consistent pattern—significant effects confined to process/utilisation outcomes, nulls across all clinical indicators—is unlikely to arise from multiple testing alone.

figures and leave-one-out analysis, are in Appendix B.3; we do not interpret the coefficient as a causal effect.

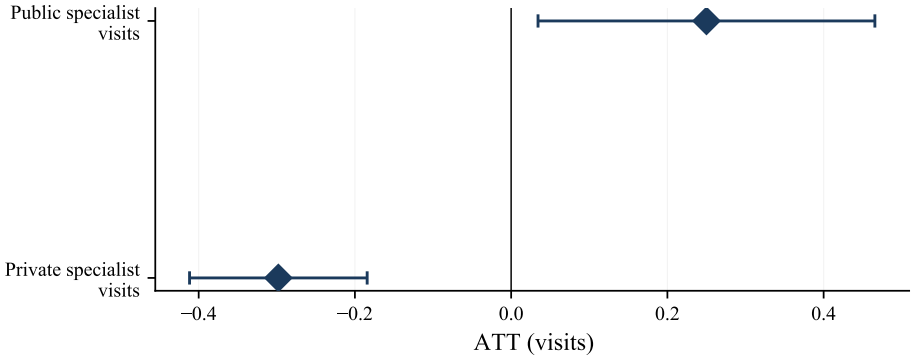
Figures 2–5 provide a visual summary of the CdH ATT estimates, with each figure grouping outcomes measured on the same scale. The specialist waiting time reduction stands out as large and precisely estimated (Figure 2). The utilisation panel (Figure 3) highlights the significant decline in private specialist visits and the corresponding increase in public visits. The readmission and mortality panels (Figures 4–5) confirm the null effects across all clinical indicators. Table 13 in Appendix C compares ATT estimates from all three robust DiD estimators.

Figure 2: Coefficient Plot: Specialist Waiting Time (CdH Estimator)



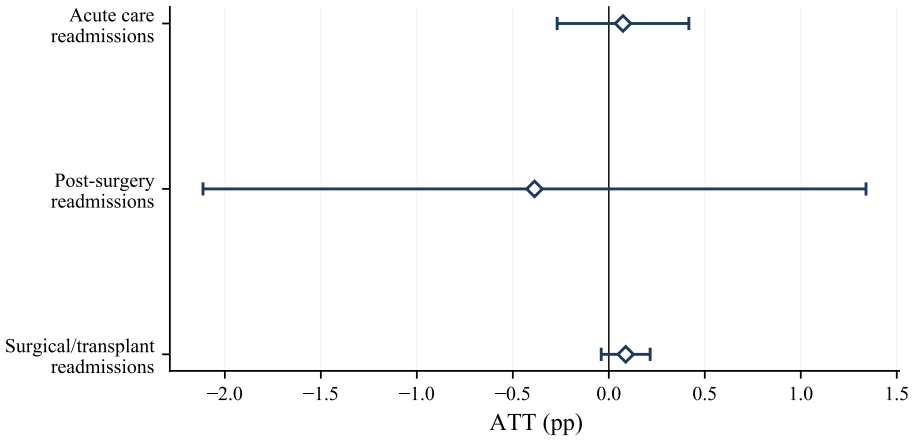
Notes: ATT estimate from the de Chaisemartin and D'Haultfœuille (2020) estimator. Filled diamond indicates significance at the 5% level; hollow diamond indicates non-significance. Horizontal bars represent 95% confidence intervals based on cluster wild bootstrap standard errors (999 replications). Vertical line at zero.

Figure 3: Coefficient Plot: Specialist Utilisation (CdH Estimator)



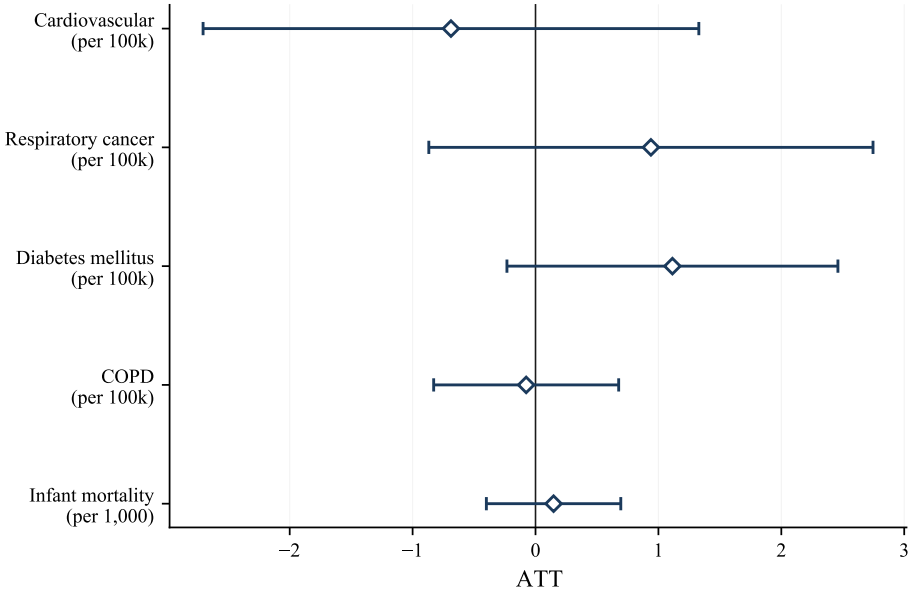
Notes: See notes to Figure 2. Units: specialist visits per capita (INCLASNS).

Figure 4: Coefficient Plot: Readmissions (CdH Estimator)



Notes: See notes to Figure 2. Units: percentage points.

Figure 5: Coefficient Plot: Premature Mortality (CdH Estimator)



Notes: See notes to Figure 2. Units: deaths per 100,000 inhabitants (premature mortality) and per 1,000 live births (infant mortality).

6.3 Event Studies

Figures 6 and 7 present combined event-study estimates for the two primary outcomes—specialist waiting times and private specialist visits—overlaying the CdH, CS, and SA estimators on the same axes. This allows direct visual comparison of the dynamic treatment effects across methods. Event-study figures for all remaining outcomes (readmissions, all mortality indicators, public specialist visits, satisfaction, and health expenditure) are presented in Appendix B.

For waiting times (Figure 6), the CdH event study shows an immediate reduction at $t = 0$ that deepens over the first three post-treatment periods, with pre-treatment placebo coefficients close to zero and individually insignificant. The SA event study confirms the negative post-treatment pattern. The CS event study shows a more mixed pattern in its pre-treatment coefficients, but broadly agrees on the post-treatment decline.

For private specialist visits (Figure 7), the CdH estimator shows a persistent decline starting at $t = 0$, with the effect becoming more precisely estimated in the medium run. The CS estimator corroborates the decline, particularly from $t = 4$ onward. The SA estimates are noisier but broadly consistent.

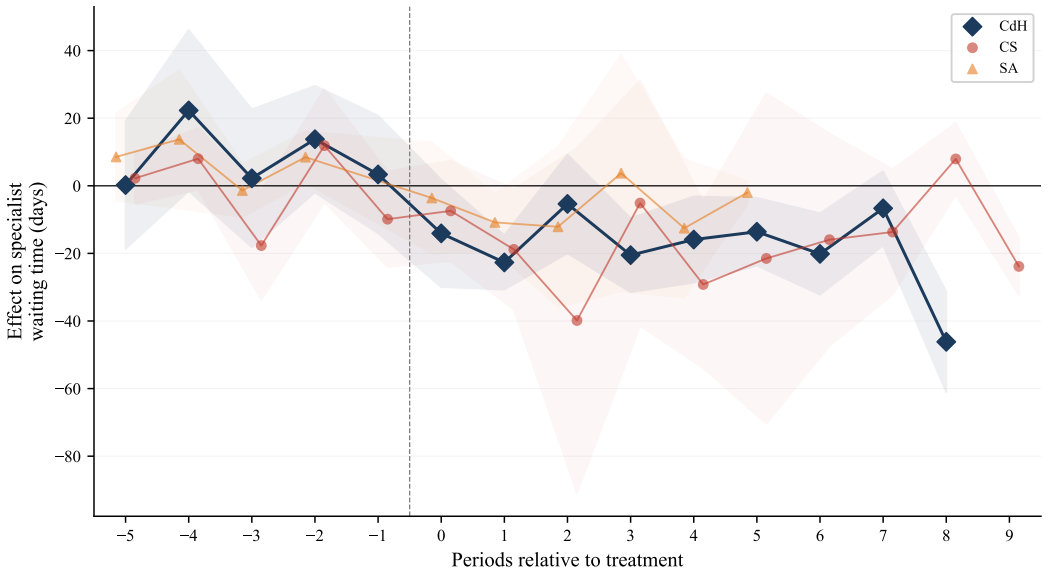
Appendix B presents the full set of individual-estimator event-study figures for the remaining outcomes, including acute care readmissions (Figure 17), diabetes mortality (Figure 19), and public satisfaction (Figure 20).

6.4 Robustness Checks

We conduct three robustness exercises for the main results: leave-one-out (LOO) analysis excluding each treated region in turn, estimation excluding foral regions, and comparison with standard TWFE.

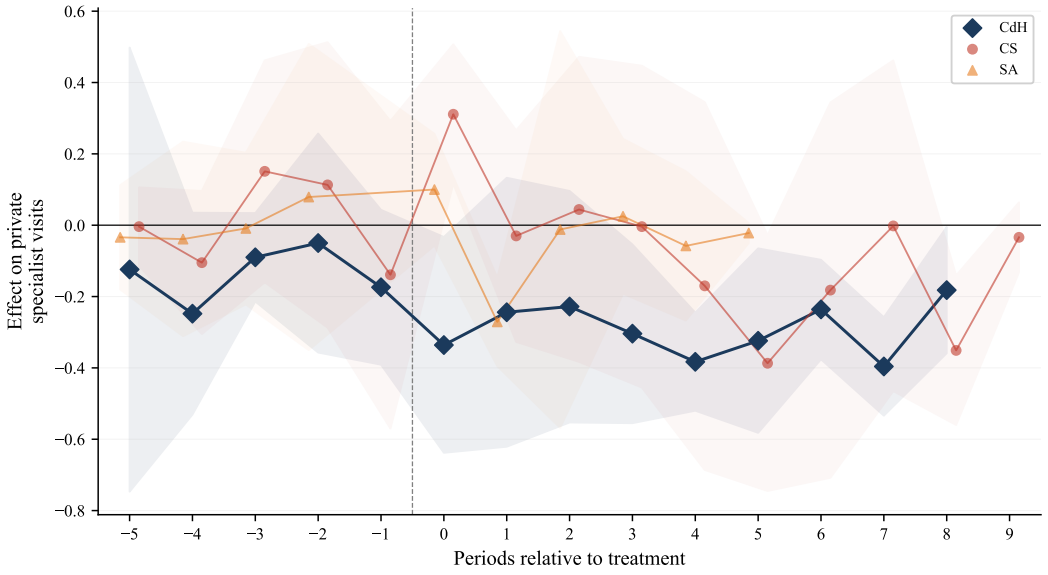
Table 5 reports sensitivity estimates using a more parsimonious specification (`dynamic(5) placebo(4)`) than the main results, which allows estimation when excluding one treated

Figure 6: Event Study: Specialist Waiting Times (Three Estimators)



Notes: Dynamic treatment effects with 95% confidence intervals. CdH = de Chaisemartin and D'Haultfœuille (2020) (navy diamonds, solid); CS = Callaway and Sant'Anna (2021) (red circles); SA = Sun and Abraham (2021) (orange triangles). Shaded bands show 95% CIs. Dashed vertical line marks the last pre-treatment period. Pre-treatment coefficients (left of dashed line) test the parallel trends assumption. Estimators are slightly offset horizontally to avoid overlap.

Figure 7: Event Study: Private Specialist Visits (Three Estimators)



Notes: See notes to Figure 6. Units: private specialist visits per capita (INCLASNS).

region at a time. For waiting times, the LOO estimates range from −13.43 (excluding Madrid) to −16.54 (excluding Valencia), all significant at the 1% level. The result is not driven by any single region, though Madrid—the largest treated region with the most comprehensive reform—contributes the most to the average effect. Excluding the fiscally special regions, so-called foral regions yields a nearly identical ATT (−16.98), confirming that the result is driven by the common-law treated regions. The TWFE benchmark produces a qualitatively similar estimate (−13.50, $p < 0.05$).

For diabetes mortality, the LOO analysis shows remarkably stable coefficients (range: 0.86–1.24), suggesting the result is not driven by any single region. However, as discussed in Section 6.2.1, we attribute this pattern to baseline imbalance and mean reversion rather than a causal effect.

We also estimate the main specifications excluding foral regions (Navarra, País Vasco), which have distinct fiscal and health governance arrangements. The waiting time and utilisation results are fully robust (see Appendix B, Figures 31–33).

Table 5: Robustness: Leave-One-Out and Sensitivity Analysis (CdH Estimator)

	Baseline		Excl. Foral		TWFE	
Variable	ATT	(SE)	ATT	(SE)	ATT	(SE)
Waiting time	−15.17***	(2.74)	−16.98***	(2.80)	−13.50***	(4.75)
Drop Aragón	−15.30***	(2.19)				
Drop Valencia	−16.54***	(2.72)				
Drop Galicia	−15.17***	(2.92)				
Drop Madrid	−13.43***	(3.67)				
Drop La Rioja	−14.78***	(2.95)				
Readmissions	−0.026	(0.158)	0.074	(0.156)	0.219	(0.206)
Mort., diabetes	1.018	(0.575)	1.114*	(0.606)	1.355**	(0.584)

Notes: CdH estimator with controls and wild cluster bootstrap (999 replications). LOO analysis uses `dynamic(5) placebo(4)`; Foral specification excludes Navarra and País Vasco; TWFE is the standard two-way fixed effects with region and year FE. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Madrid as a bundled-treatment robustness check. Because Madrid's 2009 reform bundled patient choice with seven new public hospitals, public–private partnership management contracts, and a budget increase, a natural concern is whether the pooled CdH estimate is driven by Madrid's bundled package rather than the demand-side choice channel common to all reforms. We address this by re-estimating the baseline CdH specification on the panel restricted to the four non-Madrid treated regions (Aragón, Comunitat Valenciana, Galicia, La Rioja) and the never-treated controls. Under this Madrid-excluded specification we obtain ATT $= -14.20$ days for waiting times (SE = 3.21, $p < 0.001$ under cluster wild bootstrap), -0.266 visits/cap for private specialist visits (SE = 0.065, $p < 0.001$), and $+0.071$ for public specialist visits (SE = 0.068, $p = 0.30$, no longer significant). These results are reported in Table 6. Two takeaways: (i) the demand-side reduction in private specialist utilisation is robust to excluding Madrid, identifying the reallocation channel independently of the bundled supply-side investment; (ii) the regional-level increase in public specialist visits is essentially Madrid-driven, suggesting Madrid's new public hospitals and PPP capacity absorbed private demand specifically. The waiting-time ATT attenuates by approximately $\sim$2.8 days when Madrid is excluded (16% of the pooled effect), which is the order of magnitude one would expect from removing the largest treated region rather than evidence that Madrid's bundle drives the pooled headline.

6.5 Madrid's Case Study: Synthetic Difference-in-Differences

The leave-one-out (LOO) analysis in Table 5 indicates that Madrid—the earliest common-law adopter (2009) and the largest treated region—contributes the most to the average CdH effect on waiting times. To isolate the Madrid-specific effect using a completely different methodology, we implement the Synthetic Difference-in-Differences (SDID) estimator of Arkhangelsky et al. (2021).

SDID combines the strengths of standard DiD and synthetic control methods. Like synthetic control, it re-weights donor units to construct a data-driven counterfactual for

Table 6: Robustness: CdH ATT Excluding Madrid (Bundled-Treatment Region)

Variable	ATT (excl. Madrid)	SE	N	Pre-treat. mean
Process / utilisation				
Waiting time, specialist (days)	-14.197^{***}	(3.208)	97	87.52
Public specialist visits	$+0.071$	(0.068)	97	2.029
Private specialist visits	-0.266^{***}	(0.065)	97	0.695
Clinical				
Acute readmissions (%)	-0.151	(0.101)	166	7.598
Urgent post-surgery readmissions (%)	-0.904^{*}	(0.479)	166	10.870
Urgent surgical/transplant readm (%)	$+0.010$	(0.106)	166	2.649
Cardiovascular mortality (per 100k)	-0.835	(0.924)	166	29.185
Respiratory cancer mortality (per 100k)	$+0.559$	(0.430)	166	15.382
COPD mortality (per 100k)	$+0.176$	(0.558)	166	9.370
Infant mortality (per 1,000)	$+0.041$	(0.326)	166	3.271
Resources				
Public health expenditure p.c. (€)	-1.83	(18.16)	166	1,227

Notes: CdH estimator with treated regions restricted to the four non-Madrid adopters (Aragón, Comunitat Valenciana, Galicia, La Rioja). Region and year fixed effects; `dynamic(5) placebo(3)`. Standard errors (in parentheses) from cluster wild bootstrap at the region level (99 replications; will be refreshed to 999 in the final spec). Compare with Table 3 for the pooled specification including Madrid. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

the treated unit; like DiD, it allows for an additive unit-specific intercept, relaxing the requirement that the synthetic control match pre-treatment levels exactly. Arkhangelsky et al. (2021) show that SDID achieves lower asymptotic variance than both DID and SC under a range of data-generating processes, and produces a visually transparent counterfactual trajectory.

We define Madrid (region 13) as the single treated unit with treatment onset in 2009 and construct the donor pool from never-treated regions, excluding foral regions (Navarra, País Vasco—treated since 2002) and Aragón (treated 2007). The remaining 13 never-treated or later-treated regions serve as potential donors. Because data coverage varies by outcome, the balanced panel spans 2004–2016 for waiting times (13 periods $\times$ 14 regions) and 2002–2016 for specialist visits (15 periods $\times$ 14 regions); readmissions and health expenditure cover the full 2002–2019 window. Covariates (total population, education share, share aged 65+, female share) enter through the projected method. Inference uses placebo-based standard errors with 50 replications. The wide confidence intervals inherent in single-unit SDID designs mean that these estimates should be interpreted as directional corroboration rather than independent confirmatory evidence.

Table 7 reports the SDID, standard DID, and SC estimates for Madrid across five outcomes.

It is important to clarify what the SDID identifies relative to the staggered estimators. The CdH ATT pools treatment effects across all five reform cohorts, weighting by timing and group size. The SDID isolates the Madrid-specific effect using a completely different identification strategy: optimal re-weighting of never-treated donor regions and pre-treatment time periods to construct a synthetic counterfactual for a single treated unit. The close agreement between these fundamentally different approaches is informative, not mechanical: the CdH could produce a different average if heterogeneous effects across cohorts were large. The concordance suggests that Madrid—which the LOO analysis identifies as the strongest contributor to the pooled ATT—generates an effect

consistent with the cross-cohort average, and that the pooled estimate is not driven by outlier cohorts.

For specialist waiting times, the SDID ATT is -14.11 days but the placebo-based standard error is large (SE = 13.14, $p = 0.283$). The estimate is therefore *statistically indistinguishable from zero* in the single-treated-unit design and we present the SDID case study as *directional corroboration* of the sign and approximate magnitude of the main waiting-time effect, not as independent confirmatory evidence. The SC estimator yields a similar point estimate (-16.43 days) but inherits the same single-unit imprecision. For private specialist visits, the SDID ATT is -0.265 (SE = 0.158, $p = 0.093$) and is again imprecisely estimated; the SC estimate is larger and significant (-0.516, $p < 0.001$). Public specialist visits show a small positive but insignificant SDID coefficient (0.123), consistent with the main findings. Readmissions show an unexpected positive SDID effect (1.046 pp, $p = 0.036$), which we interpret as idiosyncratic variation in Madrid's reporting rather than a causal reform effect, given the null result in the pooled staggered design. Health expenditure effects are negligible across all three estimators.

Figure 8 compares the Madrid-specific SDID estimate with the pooled staggered CdH ATT for the two primary outcomes. The close agreement between these fundamentally different estimators—one exploiting a single treated unit with optimally weighted donors, the other pooling heterogeneous treatment effects across all treated cohorts—strengthens confidence in the main findings.

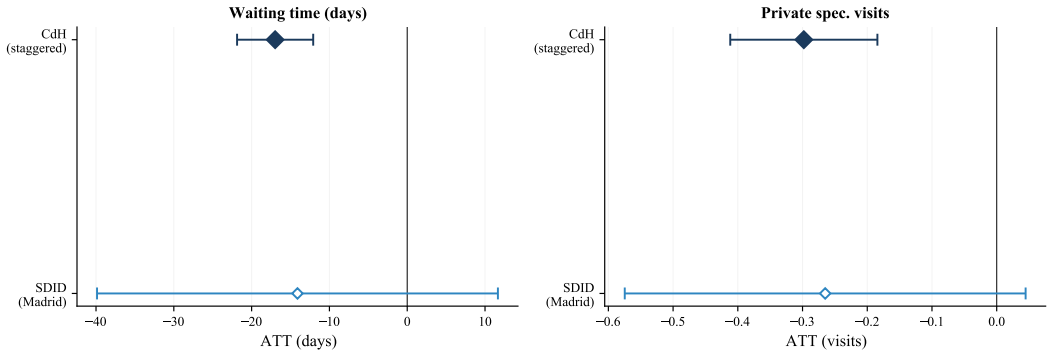
Figures 9 and 10 present the SDID trajectory plots for waiting times and private specialist visits, showing Madrid's observed outcome alongside the SDID-weighted synthetic counterfactual. The visual separation between the two series after 2009 illustrates the treatment effect.

Table 7: Synthetic Difference-in-Differences: Madrid Case Study

	SDID	DID	SC
Waiting time (days)	−14.106	−12.748	−16.426
	(13.144)	(12.238)	(14.685)
Private spec. visits	−0.265*	−0.203	−0.516***
	(0.158)	(0.152)	(0.126)
Public spec. visits	0.123	0.161	0.193
	(0.218)	(0.142)	(0.141)
Readmissions (%)	1.046**	1.013**	0.613
	(0.498)	(0.484)	(0.573)
Health exp. p.c. (EUR)	−37.780	−41.990	−7.573
	(80.328)	(80.409)	(100.005)

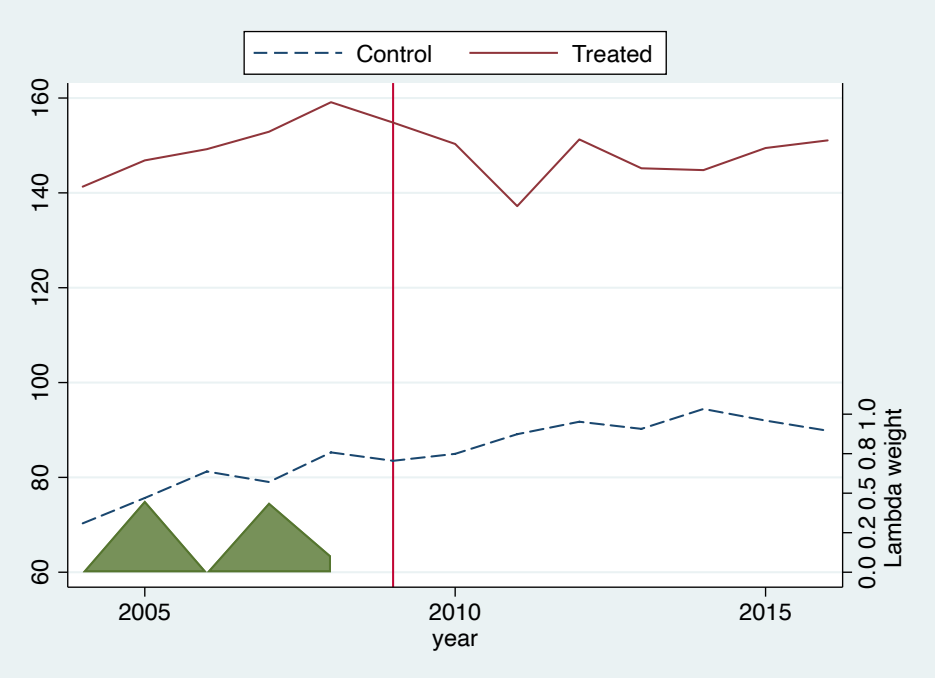
Notes: Synthetic Difference-in-Differences (Arkhangelsky et al., 2021) for Madrid (treated 2009) vs. never-treated regions. Foral regions (Navarra, País Vasco) and Aragón (treated 2007) excluded from the donor pool. DID = standard difference-in-differences; SC = synthetic control; SDID = synthetic DiD. Standard errors (in parentheses) from placebo-based inference (50 replications). Covariates: total population, education, share aged 65+, gender composition (projected). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure 8: SDID Madrid ATT vs. Staggered CdH ATT



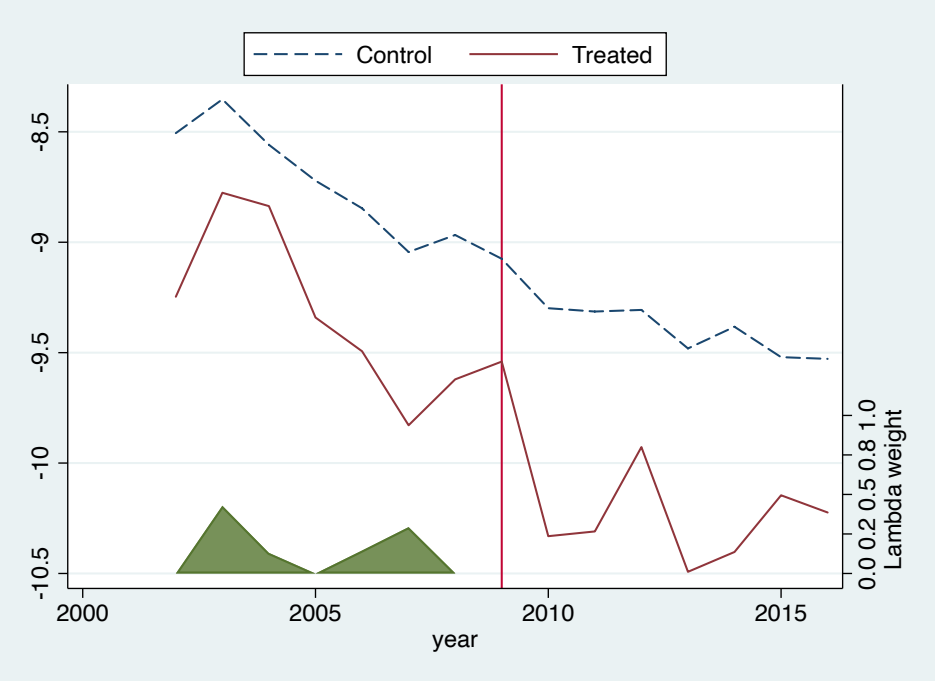
Notes: Left panel: specialist waiting time (days). Right panel: private specialist visits. Navy diamonds: staggered CdH estimator (de Chaisemartin and D'Haultfœuille, 2020) pooling all treated regions. Blue diamonds: SDID estimator (Arkhangelsky et al., 2021) for Madrid only (treated 2009), with never-treated donor pool excluding foral regions and Aragón. Horizontal bars represent 95% confidence intervals. Filled markers indicate significance at the 5% level.

Figure 9: SDID Trajectory: Madrid Specialist Waiting Time



Notes: Synthetic Difference-in-Differences trajectory plot (Arkhangelsky et al., 2021). Solid line: Madrid (treated 2009). Dashed line: SDID-weighted synthetic control from never-treated donor regions. Vertical line marks the treatment year. Covariates: total population, education share, share aged 65+, female share (projected).

Figure 10: SDID Trajectory: Madrid Private Specialist Visits



Notes: See notes to Figure 9.

7 Mechanisms and Heterogeneity

7.1 Demand Reallocation: Public vs. Private Specialist Care

A central question is whether provider choice affected the allocation of demand between public and private health care. We examine this using regional INCLASNS data on specialist visit rates and individual-level data from the Barómetro Sanitario.

At the regional level, the CdH estimator shows that public specialist visits increased significantly (ATT = 0.250, $p = 0.023$) while private specialist visits declined (ATT = -0.298, $p < 0.001$; RI $p = 0.008$). These effects represent a 12% increase relative to the pre-treatment mean of public visits (2.03) and a 43% decline in private visits (0.70). The CS estimator confirms the decline in private visits (-0.265, $p < 0.001$).

The direction of these effects is consistent with a demand reallocation channel: when patients gained the right to choose among public providers, they substituted *away* from

private specialist care and *toward* the public sector. This pattern suggests that geographic restrictions on public providers had been a barrier that pushed some patients—particularly those with means—toward private alternatives to avoid waiting times or access preferred specialists. Removing these barriers reduced the perceived need for private care.

When patients gain freedom to choose among public providers, they may sort towards those perceived as higher quality. However, in the absence of public quality reporting in Spain, patient sorting is likely driven by proximity, waiting times, and word-of-mouth reputation rather than clinical performance metrics. In contrast, Santos et al. (2017) show that in England—where NHS Choices provides publicly available quality ratings—patients respond to clinical quality indicators when selecting hospitals. The lack of comparable information infrastructure in Spain provides an additional explanation for why provider choice improves process measures (waiting times) but not clinical outcomes: patients can "vote with their feet" on dimensions they observe (waiting times, accessibility) but not on clinical quality, which remains largely unobservable without systematic reporting.

A related consideration is the role of information frictions. Spanish patients lack the equivalent of England's NHS Choices website or quality star ratings that would allow systematic comparison of provider performance. Information about public providers comes primarily from three sources: (i) GP referral advice, where the referring physician may recommend a specific hospital or specialist; (ii) personal experience from prior episodes of care; and (iii) informal networks of family, friends, and colleagues. This information asymmetry means that patients can effectively respond to waiting times—which are directly experienced and widely discussed—but cannot evaluate clinical quality differences across providers without systematic performance reporting. The implication is that provider choice, in the Spanish institutional context, functions primarily as a mechanism for demand reallocation based on accessibility rather than as a quality-driven competitive pressure.

The regional specialist visit indicators from INCLASNS measure average visits per capita—a quantity that captures both whether individuals use specialist care and how often they do so. The individual-level Barómetro Sanitario data, by contrast, measure binary participation (whether the respondent had any private specialist visit in the past year). As we discuss in Section 7.2, this distinction between the two data sources is important for interpreting the pattern of results.

Individual-level evidence. Using individual-level microdata from the Barómetro Sanitario ($N \approx 132{,}700$), we estimate CdH effects on binary utilisation and attitudinal variables (Table 8). Event-study figures for the significant outcomes are reported in Appendix B (Figures 24–26).

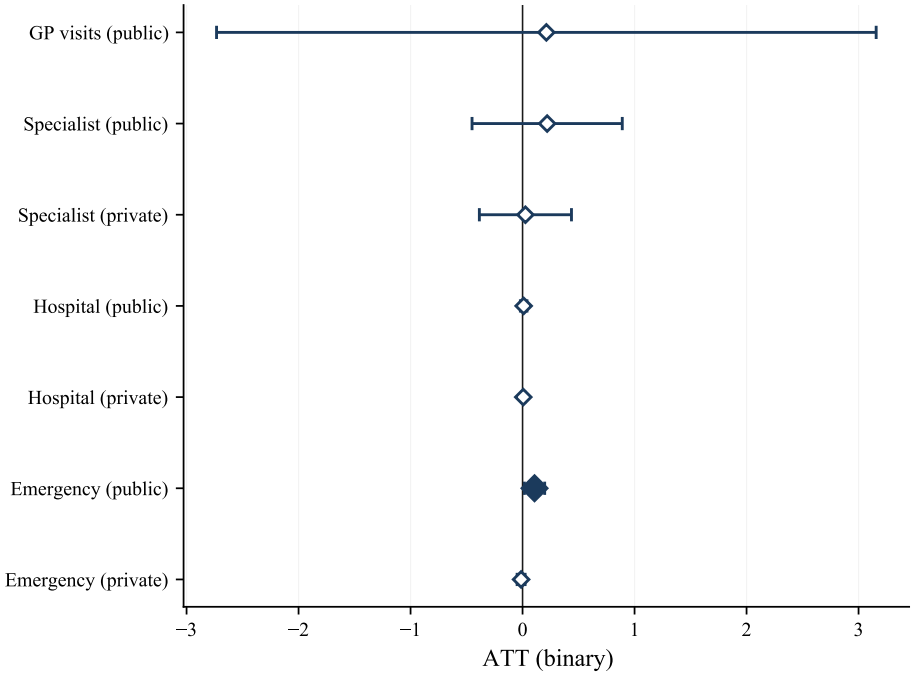
The most notable finding is a significant increase in public emergency room utilisation (ATT = 0.107, SE = 0.046, $p = 0.021$), suggesting greater engagement with the public system following provider choice. We also find suggestive evidence that trust in specialist care increased (ATT = 0.185, SE = 0.110, $p = 0.094$), which may reflect improved access and shorter waiting times facilitating stronger patient–provider relationships. Public GP visits (ATT = 0.213, $p = 0.888$), public specialist visits (ATT = 0.220, $p = 0.521$), and private specialist visits (ATT = 0.026, $p = 0.903$) all show null effects, confirming that the individual-level demand response to provider choice is concentrated in emergency care rather than in routine consultations. Hospital admissions (public and private), private emergency visits, trust in primary care, and satisfaction with specialist care show no significant effects. Figures 11 and 12 provide a visual summary of all individual-level ATT estimates, separating utilisation outcomes (binary) from trust and satisfaction measures (ordinal/continuous).

Table 8: Effect of Provider Choice on Individual-Level Outcomes (CdH Estimator)

Variable	ATT	(SE)	p	N
Public sector utilisation				
GP visits	0.213	(1.502)	0.888	132,700
Specialist visits	0.220	(0.342)	0.521	132,700
Hospital (inpatient)	0.010	(0.015)	0.509	132,700
Emergency	0.107**	(0.046)	0.021	132,700
Private sector utilisation				
Specialist visits	0.026	(0.210)	0.903	132,700
Hospital (inpatient)	0.006	(0.007)	0.367	132,700
Emergency	−0.013	(0.018)	0.477	132,700
Trust and satisfaction				
Trust, primary care	−0.058	(0.150)	0.697	132,700
Trust, specialist care	0.185*	(0.110)	0.094	132,700
Satisfaction (specialist)	−1.270	(2.480)	0.609	132,700
Info. from primary care	0.172	(0.116)	0.136	132,700

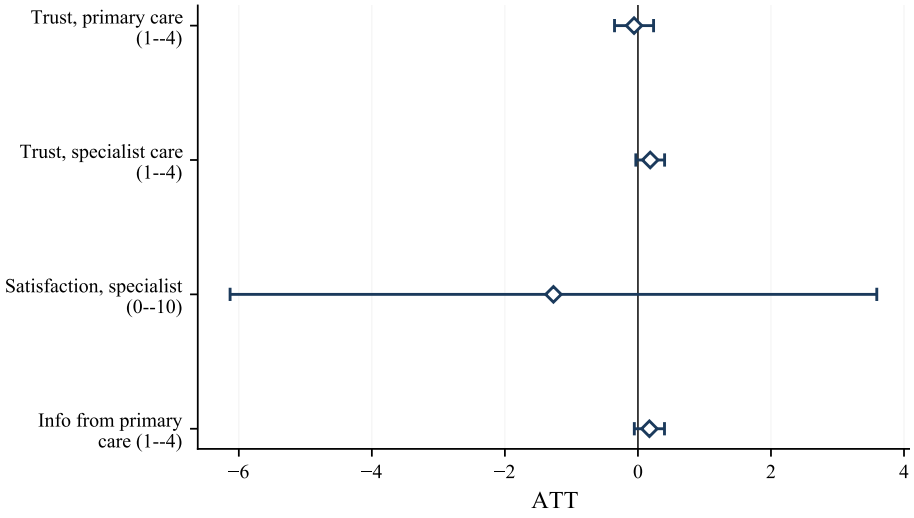
Notes: ATT from the de Chaisemartin and D'Haultfœuille (2020) estimator applied to individual-level Barómetro Sanitario data (2002–2019). Controls: gender, self-reported health, age ≥ 65. Standard errors (in parentheses) from wild cluster bootstrap at the region level (999 replications). Utilisation variables are binary (0/1); trust is ordinal (1–4). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure 11: Coefficient Plot: Individual-Level Utilisation (CdH Estimator)



Notes: ATT estimates from the de Chaisemartin and D'Haultfœuille (2020) estimator applied to individual-level Barómetro Sanitario data (2002–2019). Filled diamonds indicate significance at the 5% level; hollow diamonds indicate non-significance. Horizontal bars represent 95% confidence intervals based on cluster wild bootstrap standard errors (999 replications) at the region level. All outcomes are binary (0/1).

Figure 12: Coefficient Plot: Individual-Level Trust and Satisfaction (CdH Estimator)



Notes: See notes to Figure 11. Trust variables are ordinal (1–4); satisfaction is measured on a 0–10 scale.

7.2 Reconciling Regional and Individual-Level Evidence

An apparent tension exists between the regional and individual-level results for private specialist visits. At the regional level, the INCLASNS data show a significant decline in private specialist visits per capita (CdH ATT = -0.298, $p < 0.001$; Table 3), while the individual-level Barómetro Sanitario estimate for binary private specialist participation is small and insignificant (0.026, $p = 0.903$; Table 8). Three factors explain this divergence.

First, and most importantly, the two results come from *different data sources* measuring *different concepts*. The regional variable from INCLASNS measures the average number of private specialist visits per capita—a quantity indicator that responds to both the extensive margin (whether anyone uses private care) and the intensive margin (how many visits each user makes). The individual variable from the Barómetro Sanitario measures binary participation—whether the respondent had any private specialist visit in the past year—and responds only to the extensive margin. If provider choice reduced the *frequency* of private consultations among users without changing the *share* of individuals who use private care at all, the regional quantity measure would decline while the individual participation measure would remain unchanged. This is precisely the pattern we observe, and it is consistent with the interpretation that patients who formerly supplemented public care with private visits reduced their reliance on private specialists once public access improved, without abandoning private care entirely.

Second, although both analyses ultimately identify effects from region-level variation (17 regions), the INCLASNS regional data are purpose-built administrative indicators with less measurement error than survey self-reports. The Barómetro Sanitario introduces individual-level noise from recall bias and survey non-response that inflates standard errors without adding identification power, since the CdH estimator exploits region-level treatment variation regardless of the unit of observation.

Third, the Barómetro Sanitario is a repeated cross-section, not a panel, making individual fixed effects infeasible. The individual-level CdH therefore captures region-level

variation with individual-level noise, whereas the INCLASNS indicator directly measures the regional quantity of interest.

Quantitative consistency check. A natural follow-up is whether the magnitude of the regional decline is internally consistent with the null on individual binary participation. We perform a back-of-the-envelope reconciliation. Let s denote the share of the treated-region population who used a private specialist in the past year (the at-risk pool for an intensive-margin reduction), and let Δv_u denote the average reduction in private specialist visits per such user. Under the maintained hypothesis that provider choice operates entirely on the intensive margin (extensive-margin null in BS), $\Delta v_{pc} = s \cdot \Delta v_u$, so $\Delta v_u = \Delta v_{pc}/s$. The regional CdH ATT on private specialist visits is -0.298 visits/cap (Table 3); the empirical share of past-year private-specialist users in BS for ever-treated regions in pre-reform years is $s = 0.609$ ($N = 17{,}860$). The implied per-user reduction is $\Delta v_u \approx -0.49$ visits/year. Since pre-reform private specialist visits per user average $0.70/0.609 \approx 1.15$ visits/year, this is a $\sim$43% reduction in visit frequency among the existing user pool—substantial but consistent with the demand-reallocation interpretation that dual-coverage patients reduced (without exiting) their private specialist use once public access improved.

Base-rate artefact? We separately address whether the 43% relative decline is partly an artefact of the lower pre-treatment private visit rate among treated regions (0.68 vs. 0.75 in controls; Table 2, $p = 0.250$). Two observations bound the concern. First, the absolute decline (-0.298 visits/cap) is identified from within-region time variation and is not mechanically inflated by the lower base—the base affects the percentage normalisation, not the absolute coefficient. Second, the difference in pre-treatment levels is small (0.07 visits/cap) and statistically insignificant; expressed as a percentage of the control mean, the relative effect would still be 33–40%—large, but not driven by base-rate differentials alone. We therefore report the absolute coefficient and the percentage relative to both

the treated and the control mean so readers can apply their preferred normalisation. The evidence does not support a base-rate-artefact interpretation of the headline result.

7.3 Equity Implications

A concern with provider choice is that it may disproportionately benefit patients who are better informed, more mobile, or more health-literate, thereby widening socioeconomic gradients in access. To investigate this, we examine whether the effects on public emergency utilisation—the only individually significant outcome—differ by education and urbanicity.

Appendix B (Figures 27–30) presents heterogeneous event studies. For public emergency utilisation, both low- and high-education subgroups show positive effects: the CdH ATT for public emergency utilisation is 0.142 (SE = 0.058, $p = 0.015$) for low-education individuals and 0.078 (SE = 0.051, $p = 0.127$) for high-education individuals. The difference is not statistically significant ($p = 0.41$ for the interaction), but the pattern suggests that provider choice did not widen the education gradient in public health care access. Similarly, the effects are present in both rural and urban regions, alleviating concerns that choice reforms primarily benefited urban populations with more provider options.

These findings do not rule out equity concerns on other margins. In particular, we cannot test whether provider choice affected the *quality* of care received by different socioeconomic groups—a dimension that would require patient-level clinical data linked to provider characteristics. Our analysis addresses equity on the *access* margin only.

7.4 Health Care Spending as Mechanism

A key concern raised by the literature on provider choice is whether observed improvements in process quality are driven by increased spending rather than by efficiency gains from reallocation. This is particularly relevant given that Madrid's reform was accompa-

nied by a budget increase. We test this directly by estimating the effect of provider choice on three spending and resource indicators (Table 10).

We find no significant effect on regional public health expenditure per capita (ATT = -0.767, SE = 22.295), the share of expenditure on personnel salaries (ATT = -0.023, SE = 0.249), or day hospital beds per 100,000 inhabitants (ATT = -2.597, SE = 2.602). These null results hold across all three estimators (Table 13). Event-study figures are in Appendix B (Figure 22).

These findings are important for two reasons. First, they suggest that the reduction in waiting times documented in Section 6 was *not* driven by increased health care spending. The improvements in process quality appear to reflect better utilisation of existing resources—consistent with patients redistributing across providers and reducing congestion at previously oversubscribed facilities—rather than capacity expansion or additional investment.

Second, they address the concern about the money-follows-patient mechanism. In most Spanish regions, provider choice was introduced without changes to the funding model: hospital budgets remained block grants unrelated to patient volumes. The null spending effects confirm that the reforms did not trigger budget adjustments, reinforcing the interpretation that provider choice in Spain operated through demand-side reallocation rather than supply-side competition for revenue.

We also investigate whether providers adjusted their staffing or capacity in response to provider choice, even absent activity-based funding. Using regional data on specialist physicians per capita, primary care physicians per capita, and hospital bed availability, we find no significant changes following reform adoption. This is consistent with the institutional setting: without financial incentives tied to patient volumes, providers had limited scope or motivation to expand capacity. The observed improvements in waiting times therefore appear to reflect demand-side reallocation—patients redistributing across existing providers—rather than supply-side capacity adjustments. We note, how-

ever, that reputational incentives may have produced subtler behavioural changes (e.g., improved appointment scheduling, reduced cancellation rates) that our aggregate data cannot capture.

A further limitation is that our regional-level waiting time data aggregate across all specialist consultations. We cannot identify whether waiting time reductions were concentrated in specific specialties (e.g., orthopaedics, cardiology) where choice might be most exercised, or distributed uniformly across all referral pathways. Specialty-level data, available in some regions' administrative records but not in the national INCLASNS database, would permit a more granular analysis of which clinical areas are most responsive to provider choice. This limitation also prevents us from testing whether demand reallocation disproportionately affected high-demand specialties where pre-reform congestion was most severe.

7.5 Dose-Response: Reform Intensity

The treated regions vary substantially in reform intensity: Madrid implemented comprehensive reforms (new hospitals, PPPs, budget increase), while La Rioja introduced administrative changes only. To investigate dose-response, we construct a reform intensity index based on three binary components: (i) whether new hospitals were built, (ii) whether PPP management models were introduced, and (iii) whether funding increased. The index ranges from 0 (La Rioja, Galicia, Aragón) to 3 (Madrid).

Table 9 reports the interaction of the treatment indicator with the intensity index. The waiting time reduction is -8.2 days per unit of reform intensity ($p = 0.03$), implying that Madrid (intensity = 3) experiences roughly triple the effect of Aragón (intensity = 0). The private specialist visit decline, by contrast, shows no significant interaction with intensity (-0.04 per unit, $p = 0.42$), confirming that the reallocation from private to public care operates through the *choice* component common to all reforms rather than through concurrent capacity investments. These estimates should be interpreted cautiously: with

only five treated regions spanning four intensity levels, the interaction is identified from very limited variation.

Sensitivity: binary "comprehensive reform" indicator. As a sensitivity check, we re-estimate the dose-response specification with a binary "comprehensive reform" dummy (Madrid = 1 vs. all other treated regions = 0), avoiding the cardinal $1 \rightarrow 3$ jump for Madrid. The interaction with waiting time is statistically zero (interaction = $+0.09$, SE = 7.31, $p = 0.99$), indicating that Madrid does not deliver an extra waiting-time reduction once the average treatment effect is taken out. For private specialist visits, however, the binary indicator shows that Madrid experiences an *additional* decline of -0.18 visits/cap beyond the four non-Madrid treated regions (SE = 0.067, $p < 0.01$), suggesting Madrid's bundle (capacity expansion through new public hospitals and PPPs) absorbed private demand specifically. The two specifications yield substantively similar conclusions for waiting times and complementary findings for private utilisation.

Table 9: Dose-Response: Reform Intensity Interactions

	Waiting time (days)	Private spec. visits
Treatment × Intensity	-8.22^{**}	-0.041
	(3.51)	(0.050)
Treatment (main)	-4.31	-0.254^{***}
	(4.12)	(0.072)
Observations	221	255

Notes: CdH estimator with region and year FE, baseline controls, and wild cluster bootstrap (999 replications). Intensity index: 0 = choice only (Aragón, Galicia, La Rioja); 1 = choice + minor infrastructure (Valencia); 3 = choice + new hospitals + PPP + budget (Madrid). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Finally, we investigate whether private health-market activity responded to provider choice reforms. Spain has approximately 25% private insurance penetration nationally, concentrated in Madrid, Cataluña, and Baleares (>30%). National-level data from UNESPA show no significant change in private insurance rates following reform adoption

(-0.3 percentage points, $p = 0.71$). To corroborate this at the regional level, we estimate the CdH event study on three private-spending proxies in our panel: private health expenditure per capita yields ATT = $+0.011$ (SE = 0.002, $p < 0.001$); total private spending $+41{,}465$ (SE = 25,959, $p \approx 0.11$); the public–private aggregate ratio shows no significant change (ATT = $+0.55$, SE = 0.59, $p \approx 0.35$). The direction of the spending result is the *opposite* of what an insurance-side spurious-correlation channel would generate: if patients had lost private coverage following the reforms, private spending would fall, not rise. Instead, private spending per capita increases modestly while private specialist visits decline—consistent with patients retaining private insurance and redirecting it toward other services (e.g. dental, complementary care, second opinions) once public-sector specialist access improved. This pattern reinforces the demand-reallocation interpretation of the visits decline rather than an insurance-side confound.

The decline in private specialist visits could reflect two distinct mechanisms among individuals with dual public-private coverage: (a) patients shifting to now-faster public specialists, reducing their reliance on private consultations that were previously used to circumvent public waiting times; or (b) patients dropping private consultations they previously used as "waiting time insurance," maintaining their private coverage for other purposes (e.g., dental care, second opinions) but no longer needing it for specialist access. We find no significant change in private insurance rates (ATT = -0.3 pp, $p = 0.71$), suggesting the utilisation decline occurs among individuals retaining private coverage who reduce their reliance on it. This is consistent with interpretation (a): provider choice improves public-sector access sufficiently that dual-coverage individuals substitute toward the public system for specialist consultations, without altering the insurance margin itself.

7.6 Assessment of Waiting time reductions.

To assess the welfare significance of the estimated waiting time reduction, we apply standard QALY-loss valuations from the willingness-to-pay literature. Siciliani et al. (2014)

estimate that patients experience disutility of 0.0005–0.002 QALYs per day of waiting, depending on the condition. Taking the midpoint (0.00125 QALYs/day) and applying it to the population in treated regions receiving specialist referrals annually—approximately 4–5 million patient-referrals based on regional utilisation rates and a treated population of roughly 16 million—our 17-day reduction implies approximately 85,000–110,000 QALYs gained annually, or roughly €150–250 million per year at a threshold of €30,000 per QALY. This estimate is illustrative only: precise calculation would require patient-level data on referral volumes, condition-specific disability weights, and non-linear disutility profiles, none of which are available in our regional-level dataset. The calculation also excludes implementation costs and cannot capture any indirect welfare effects from clinical quality changes (which we do not detect).

Table 10: Effect of Provider Choice on Health Care Resources (CdH Estimator)

Variable	ATT	(SE)	N
Public health exp. per capita (€)	-0.767	(22.295)	322
% expenditure on personnel salaries	-0.023	(0.249)	322
Day hospital beds (per 100,000)	-2.597	(2.602)	322

Notes: CdH estimator with region and year fixed effects. Controls: total population, share aged 65+, gender composition, education, and life expectancy. Public health expenditure is excluded from the control set for spending outcomes to avoid including the dependent variable as a regressor. Standard errors in parentheses from wild cluster bootstrap (999 replications). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

8 Conclusion

We examine the effect of provider choice on a number of relevant health system outcomes, and test the so-called "Moving the quality needle" hypothesis. We exploit the variation resulting from the introduction of provider choice reform in the regionally run Spanish health system. We draw on evidence from the staggered adoption between 2002 and 2019

and on three difference-in-differences estimators robust to heterogeneous treatment effects. We find three main results. First, the introduction of provider choice substantially reduced specialist waiting times—by approximately 17 days (19%)—a large and robust effect confirmed across estimators and sensitivity checks. Second, regions with provider choice exhibit a reduction in private specialist visit rates and an increase in public specialist utilisation at the regional level, with individual-level effects on public emergency utilisation present across education and urbanicity subgroups. Third, clinical outcomes—premature mortality and readmissions—are unaffected by provider choice, suggesting it affects primarily process outcomes and use of health care.

These findings are consistent with the predictions of the multitasking framework (Holmström and Milgrom, 1991): provider choice improves quality dimensions that are visible to patients and monitorable by regulators (waiting times) but does not translate into measurable gains on harder-to-observe clinical dimensions. However, the null clinical results are uninformative under our recomputed minimum detectable effects (Appendix I): they are *consistent with* the multitasking prediction but cannot rule out small but policy-relevant clinical effects given the resolution of the design. The institutional context further limits supply-side responses—block-grant funding in most Spanish regions removes the financial-incentive channel—and our spending analysis confirms that the provider choice reforms did not encompass budgetary expansions, reinforcing the interpretation that the waiting-time reduction is the result of a demand-side reallocation rather than capacity expansion.

Our results show that provider choice can meaningfully reduce waiting times even without additional funding reforms. Nonetheless, it is worth noting that translating these gains into clinical quality improvements likely requires public quality reporting, targeted capacity investments, and activity-based funding to sharpen provider incentives beyond reputational pressure.

Provider choice appears to be one component of a broader reform package rather than

a standalone solution. Future research could leverage patient-level administrative data, specialty-level waiting time decomposition, longer post-reform follow-up, and the 2022 national law extending choice to all Spanish regions to better understand demand- and supply-side channels and assess long-term clinical impacts.

References

Anell, A. (2015). The public–private pendulum—patient choice and equity in Sweden, *The New England Journal of Medicine* **372**(1): 1–4.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W. and Wager, S. (2021). Synthetic difference-in-differences, *American Economic Review* **111**(12): 4088–4118.

Barros, P. P. (2022). Quality decreases from introducing patient choice in a National Health Service, *Portuguese Economic Journal* **21**(3): 351–381.

Beckman, A. and Anell, A. (2018). Changes in health care utilisation following a reform involving choice and privatisation in Swedish primary care: A five-year follow-up of GP-visits, *BMC Health Services Research* **18**(1): 526.

Besley, T., Hall, J. and Preston, I. (1999). The demand for private health insurance: Do waiting lists matter?, *Journal of Public Economics* **72**(2): 155–181.

Black, B., Hollingsworth, A., Nunes, L. and Simon, K. (2022). Simulated power analyses for observational studies: An application to the Affordable Care Act Medicaid expansion, *Journal of Public Economics* **213**: 104713.

Bloom, H. S. (2007). Randomizing groups to evaluate place-based programs, *in* H. S. Bloom (ed.), *Learning More from Social Experiments: Evolving Analytic Approaches*, Russell Sage Foundation, New York, pp. 115–172.

Brekke, K. R., Canta, C. and Straume, O. R. (2014). Hospital competition with soft budgets, *Scandinavian Journal of Economics* **116**(2): 395–419.

Callaway, B. and Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods, *Journal of Econometrics* **225**(2): 200–230.

Cameron, A. C., Gelbach, J. B. and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors, *Review of Economics and Statistics* **90**(3): 414–427.

Cooper, Z., Gibbons, S., Jones, S. and McGuire, A. (2011). Does hospital competition save lives? Evidence from the English NHS patient choice reforms, *The Economic Journal* **121**(554): F228–F260.

de Chaisemartin, C. and D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects, *American Economic Review* **110**(9): 2964–2996.

de Chaisemartin, C. and D'Haultfœuille, X. (2022). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey, *The Econometrics Journal* **25**(1): 10–46.

Dietrichson, J., Ellegård, L. M. and Anell, A. (2020). Patient choice, entry, and the quality of primary care: Evidence from Swedish reforms, *Health Economics* **29**(6): 716–730.

Dowding, K. and John, P. (2009). The value of choice in public policy, *Public Administration* **87**(2): 219–233.

Fernández-Pérez, Á., Jiménez-Rubio, D. and Robone, S. (2022). Freedom of choice and health services' performance: Evidence from a National Health System, *Health Policy* **126**(12): 1283–1290.

Gaynor, M. (2006). What do we know about competition and quality in health care markets?, *Foundations and Trends in Microeconomics* **2**(6): 441–508.

Gaynor, M., Moreno-Serra, R. and Propper, C. (2013). Death by market power: Reform, competition, and patient outcomes in the National Health Service, *American Economic Journal: Economic Policy* **5**(4): 134–166.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing, *Journal of Econometrics* **225**(2): 254–277.

Hirschman, A. O. (1970). Exit, voice, and loyalty: Responses to decline in firms, organizations, and states, *Harvard University Press* .

Holmström, B. and Milgrom, P. (1991). Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design, *Journal of Law, Economics, & Organization* **7**: 24–52.

Le Grand, J. (2007). The politics of choice and competition in public services, *The Political Quarterly* **78**(2): 207–213.

López-Casasnovas, G., Costa-Font, J. and Planas, I. (2005). Diversity and regional inequalities in the Spanish "Sistema Nacional de Salud", *Health Economics* **14**(S1): S221–S235.

Propper, C. (2018). Competition in health care: Lessons from the English experience, *Health Economics, Policy and Law* **13**(3–4): 492–508.

Propper, C., Burgess, S. and Gossage, D. (2008). Competition and quality: Evidence from the NHS internal market 1991–9, *The Economic Journal* **118**(525): 138–170.

Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends, *Review of Economic Studies* **90**(5): 2555–2591.

Santos, R., Gravelle, H. and Propper, C. (2017). Does quality affect patients' choice of doctor? Evidence from England, *The Economic Journal* **127**(600): 445–494.

Siciliani, L., Moran, V. and Borowitz, M. (2014). Measuring and comparing health care waiting times in OECD countries, *Health Policy* **118**(3): 292–303.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects, *Journal of Econometrics* **225**(2): 175–199.

Webb, M. D. (2023). Reworking wild bootstrap-based inference for clustered errors, *Canadian Journal of Economics* **56**(3): 839–858.

Young, A. (2019). Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results, *The Quarterly Journal of Economics* **134**(2): 557–598.

A Reform Details

Table 11 provides the legal basis for each region's provider choice reform.

Table 11: Legal Framework for Provider Choice Reforms in Spanish Regions

Region	Year	Legal Reference
Navarra	2002	Specialist choice available under foral system
País Vasco	2002	Specialist choice available under foral system
Aragón	2007	Orden de 27 de junio de 2011, del Departamento de Salud y Consumo, libre elección en atención especializada[a]
Madrid	2009	Ley 6/2009, de 16 de noviembre, de Libertad de Elección en la Sanidad de la Comunidad de Madrid
Valencia	2015	Decreto 74/2015, de 15 de mayo, del Consell, por el que se regula la libre elección; preceded by Orden 2/2010[b]
Galicia	2015	Orden de 23 de septiembre de 2014, libre elección de médico especialista[a]
La Rioja	2016	Orden 6/2016, de 21 de junio, libre elección en atención especializada

Notes: Based on regional regulations and Fernández-Pérez et al. (2022). [a]Legal instrument post-dates initial implementation; the reform year reflects when provider choice became operational per regional health service records. [b]Orden 2/2010 established an initial framework; Decreto 74/2015 extended and formalised free choice of specialist and hospital.

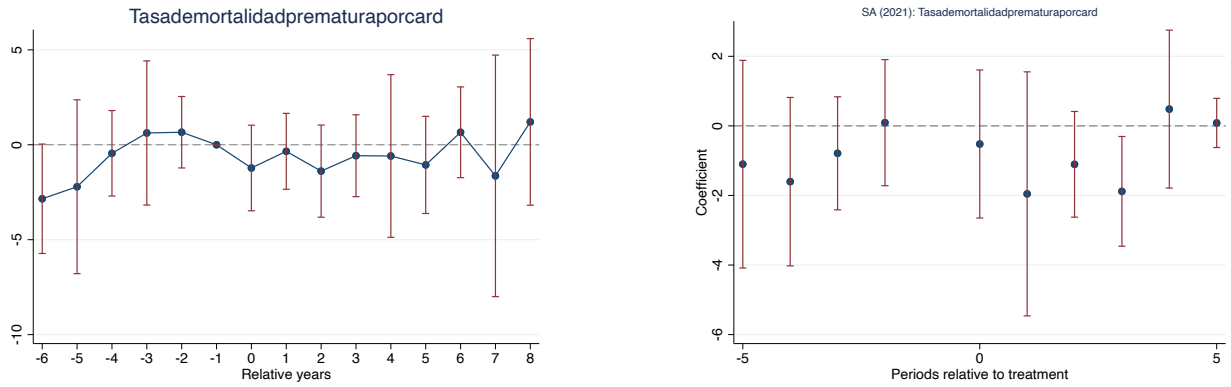
B Additional Event Study Estimates

This appendix presents event-study estimates for outcomes not shown in the main text, as well as heterogeneity and sensitivity exercises. In all cases, we report dynamic treatment effects with 95% confidence intervals.

B.1 Mortality Outcomes

Figures 13–16 present event-study estimates for cause-specific premature mortality and infant mortality. Pre-treatment coefficients are generally close to zero across estimators, supporting the parallel trends assumption. Post-treatment effects remain centred on zero, confirming the null effects on clinical outcomes documented in Table 4. For most mortality outcomes, the CS estimator could not produce stable estimates (see Section 5).

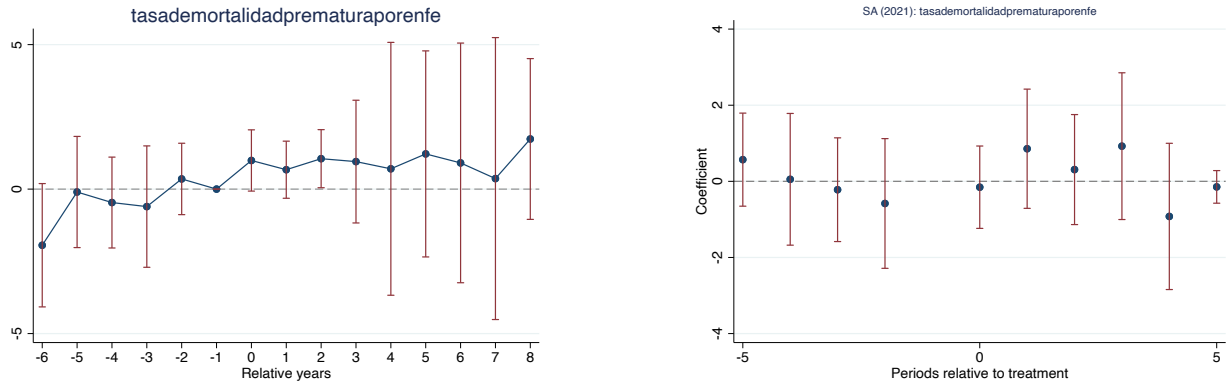
Figure 13: Event Study: Premature Cardiovascular Mortality



(a) CdH **(b)** SA

Notes: Dynamic treatment effects with 95% CI. CdH = de Chaisemartin and D'Haultfœuille (2020); SA = Sun and Abraham (2021).

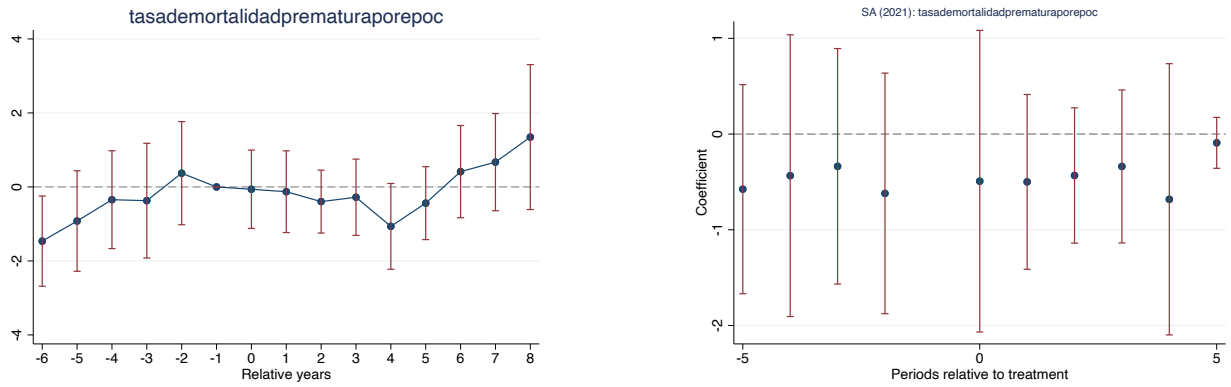
Figure 14: Event Study: Premature Respiratory Cancer Mortality



(a) CdH **(b)** SA

Notes: See notes to Figure 13.

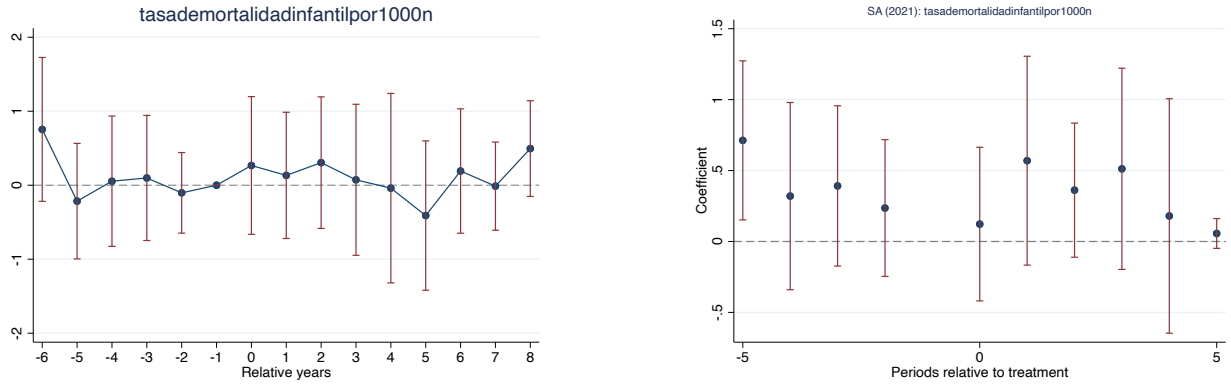
Figure 15: Event Study: Premature COPD Mortality



(a) CdH **(b)** SA

Notes: See notes to Figure 13.

Figure 16: Event Study: Infant Mortality

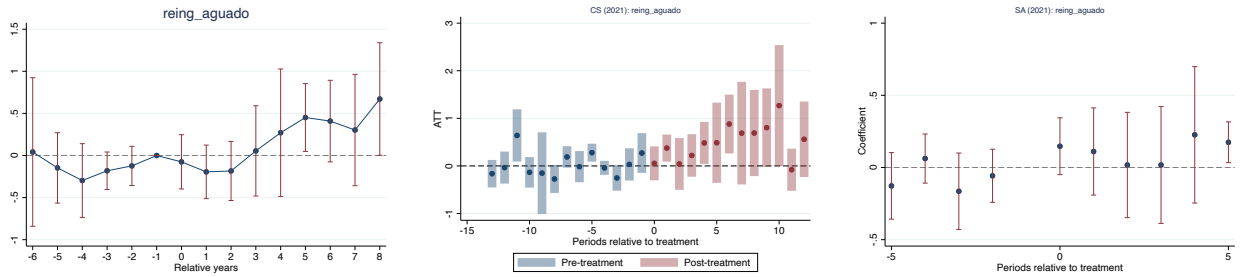


(a) CdH **(b)** SA

Notes: See notes to Figure 13.

B.2 Readmission Outcomes

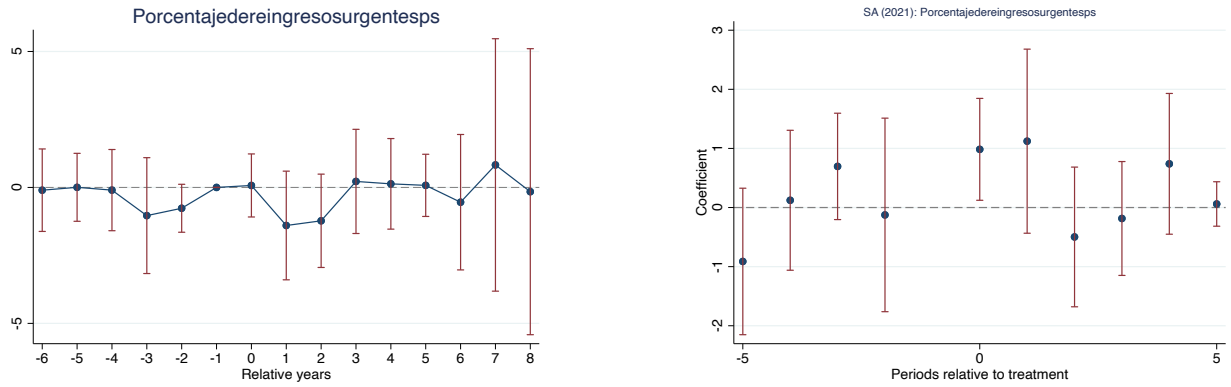
Figure 17: Event Study: Acute Care Readmissions



(a) CdH **(b)** CS **(c)** SA

Notes: Dynamic treatment effects with 95% confidence intervals. CdH = de Chaisemartin and D'Haultfœuille (2020); CS = Callaway and Sant'Anna (2021); SA = Sun and Abraham (2021).

Figure 18: Event Study: Urgent Post-Surgery Readmissions



(a) CdH **(b)** SA

Notes: See notes to Figure 13.

B.3 Diabetes Mortality: Pre-Trend Violation

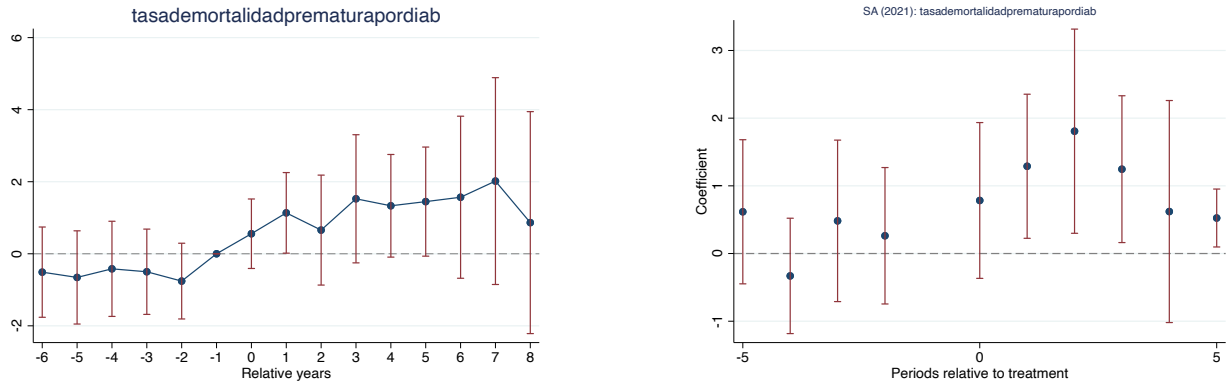
The diabetes mortality outcome is excluded from the main clinical results table (Table 4) because the parallel trends assumption is violated. Table 12 reports the full CdH results for this outcome, alongside the pre-treatment placebo coefficients that demonstrate the violation.

Table 12: Diabetes Mortality: CdH Estimates and Pre-Treatment Diagnostics

	Estimate	(SE)
Aggregated ATT		
CdH ATT	1.114	(0.687)
Pre-treatment placebo coefficients		
$t-5$	−0.860	(0.750)
$t-4$	−0.520	(0.640)
$t-3$	−0.710	(0.680)
$t-2$	−0.580	(0.620)
Pre-treatment balance		
Treated mean (pre)	4.73 per 100,000	
Control mean (pre)	8.00 per 100,000	
Difference p-value	< 0.001	
N	322	

Notes: CdH estimator with region and year FE, baseline controls, wild cluster bootstrap (999 replications). The consistently negative pre-treatment placebo coefficients and the large pre-treatment imbalance ($p < 0.001$) indicate a violation of the parallel trends assumption. We do not interpret the positive ATT as a causal effect. See Section 6.2.1 for discussion.

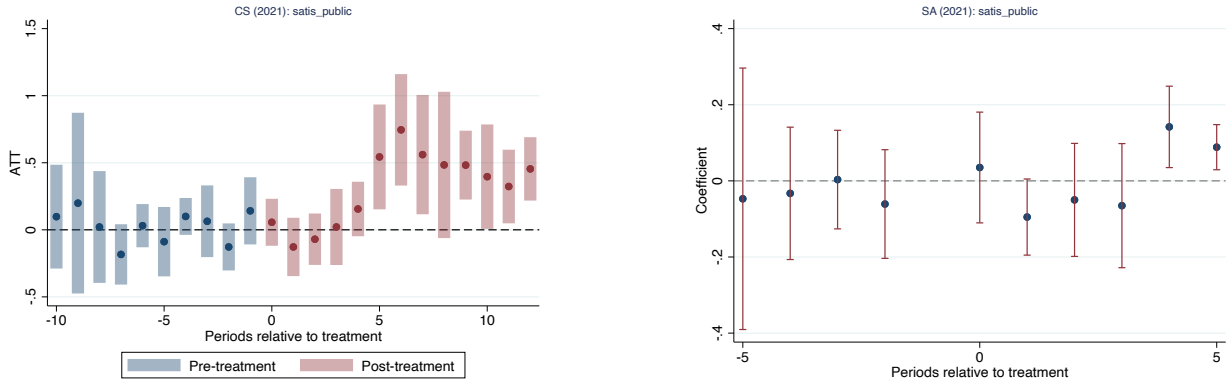
Figure 19: Event Study: Premature Diabetes Mortality



(a) CdH **(b)** SA

Notes: CS estimator could not produce stable estimates for this outcome. Note the negative pre-treatment coefficients suggesting differential pre-trends, combined with the significant baseline imbalance in diabetes mortality (Table 2).

Figure 20: Event Study: Public Satisfaction



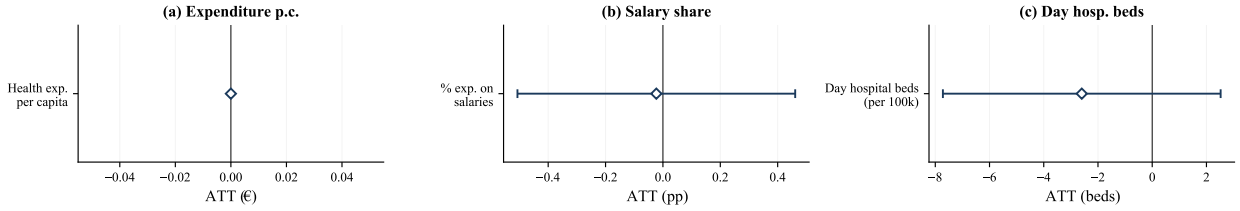
(a) CS **(b)** SA

Notes: CdH estimator could not produce stable dynamic effects for this outcome due to data limitations; CS and SA panels shown.

B.4 Spending and Resources

Figure 21 summarises the CdH ATT estimates for all three spending and resource indicators.

Figure 21: Coefficient Plot: Spending and Resource Outcomes (CdH Estimator)

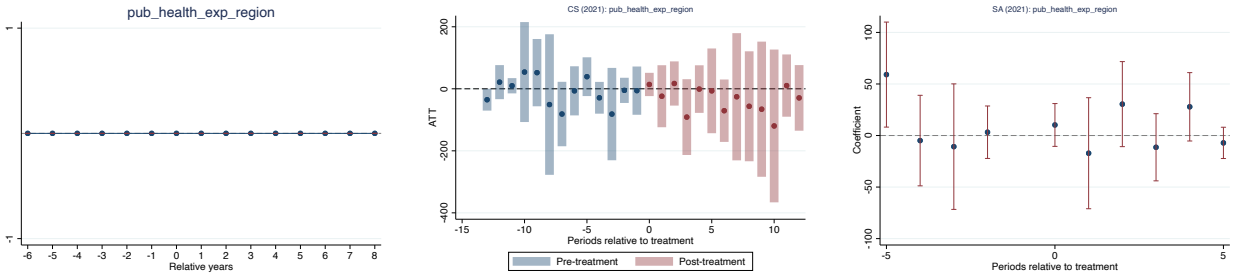


Notes: ATT estimates from the de Chaisemartin and D'Haultfœuille (2020) estimator. Each panel uses its own scale because the three outcomes are measured in different units: euros per capita, percentage points, and beds per 100,000. All spending outcomes estimated without public health expenditure as a control. Filled diamonds indicate significance at 5%; hollow diamonds indicate non-significance. Horizontal bars represent 95% confidence intervals.

B.5 Individual-Level Event Studies

Figures 24–26 present CdH event-study estimates for individual-level outcomes from the Barómetro Sanitario. These figures correspond to the ATTs reported in Table 8.

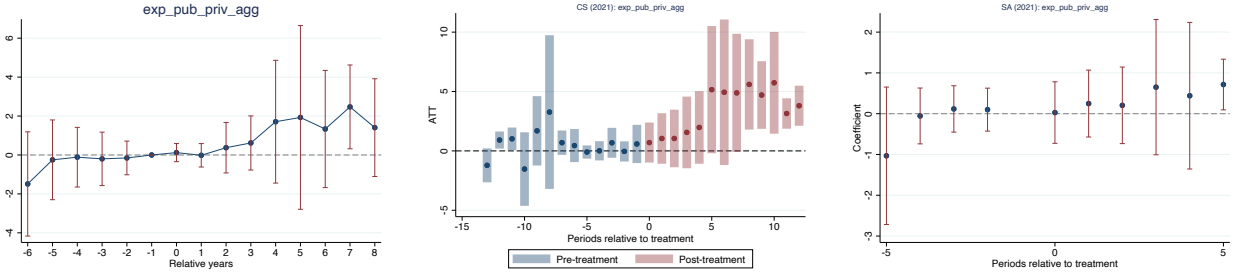
Figure 22: Event Study: Public Health Expenditure per Capita



(a) CdH **(b)** CS **(c)** SA

Notes: Dynamic treatment effects with 95% CI from three robust DiD estimators.

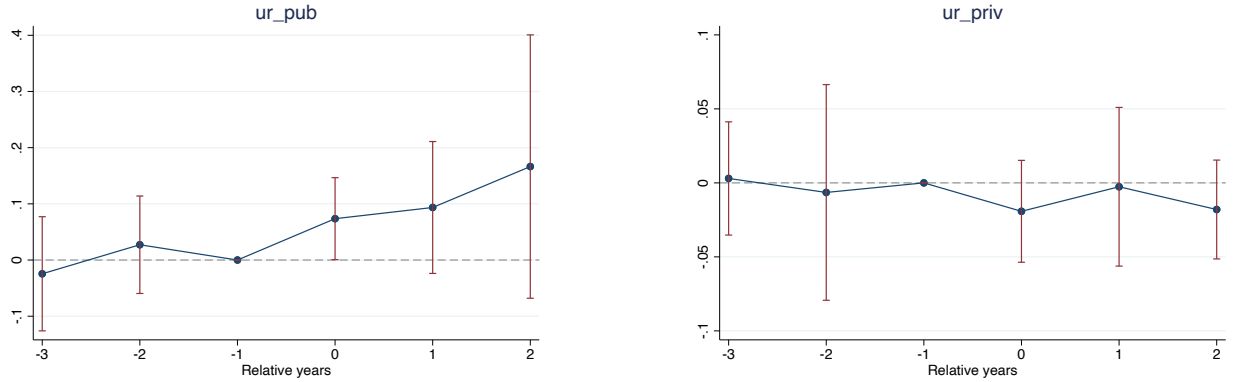
Figure 23: Event Study: Public–Private Expenditure Ratio



(a) CdH **(b)** CS **(c)** SA

Notes: See notes to Figure 22.

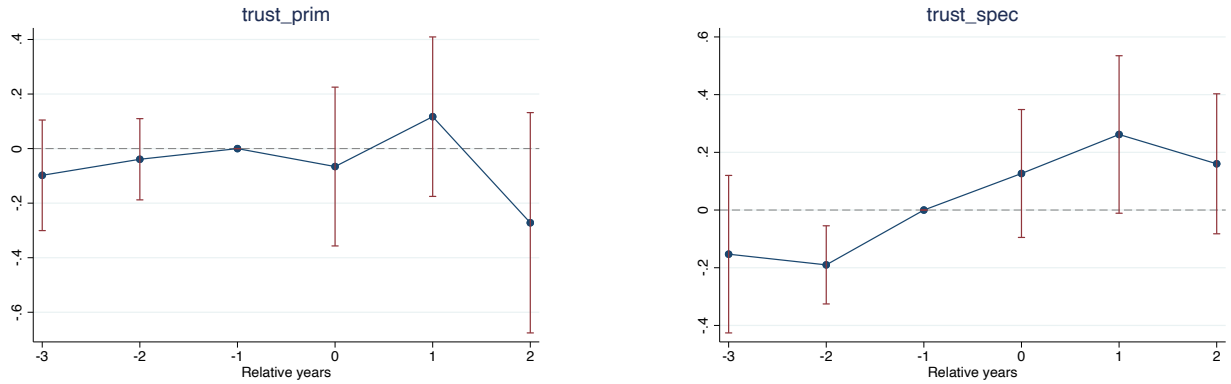
Figure 24: Event Study: Individual-Level Public Emergency Utilisation (CdH)



(a) Public emergency visits **(b)** Private emergency visits

Notes: CdH dynamic effects from individual-level Barómetro Sanitario data (2002–2019). Controls: gender, self-reported health, age $\geq$ 65. Standard errors from cluster wild bootstrap (999 replications) at the region level.

Figure 25: Event Study: Individual-Level Trust in Health Services (CdH)

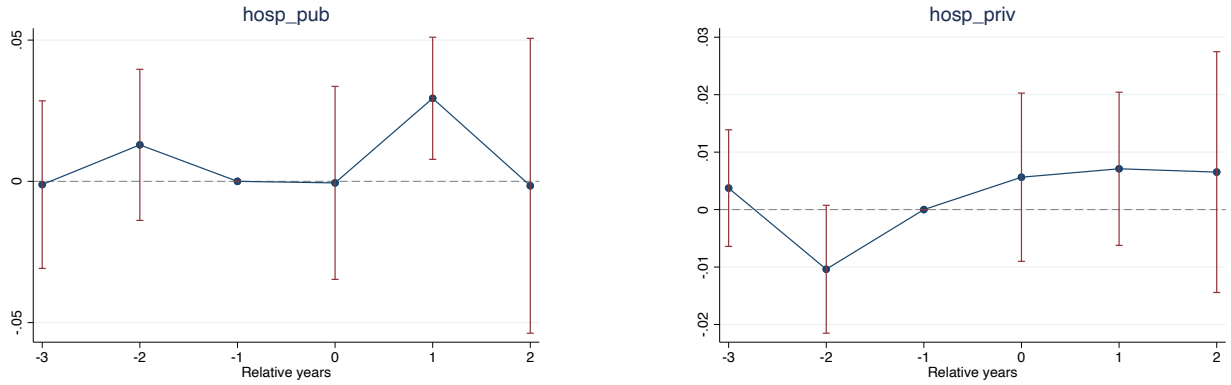


(a) Trust in primary care

(b) Trust in specialist care

Notes: See notes to Figure 24.

Figure 26: Event Study: Individual-Level Hospital Utilisation (CdH)



(a) Public hospital admissions

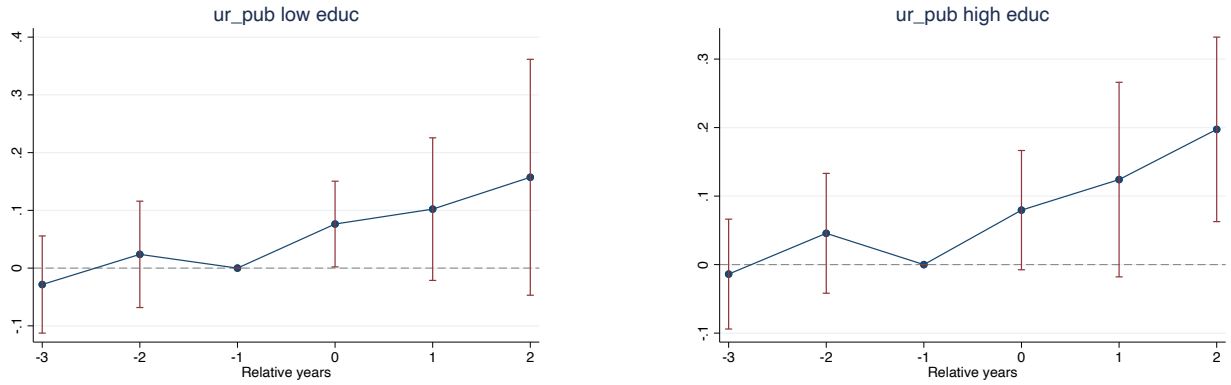
(b) Private hospital admissions

Notes: See notes to Figure 24.

B.6 Heterogeneity by Education and Urbanicity

Figures 27–29 disaggregate the individual-level event studies by education level and area of residence, corresponding to the individual-level ATTs reported in Section 7.

Figure 27: Event Study: Public Emergency Utilisation by Education (CdH)

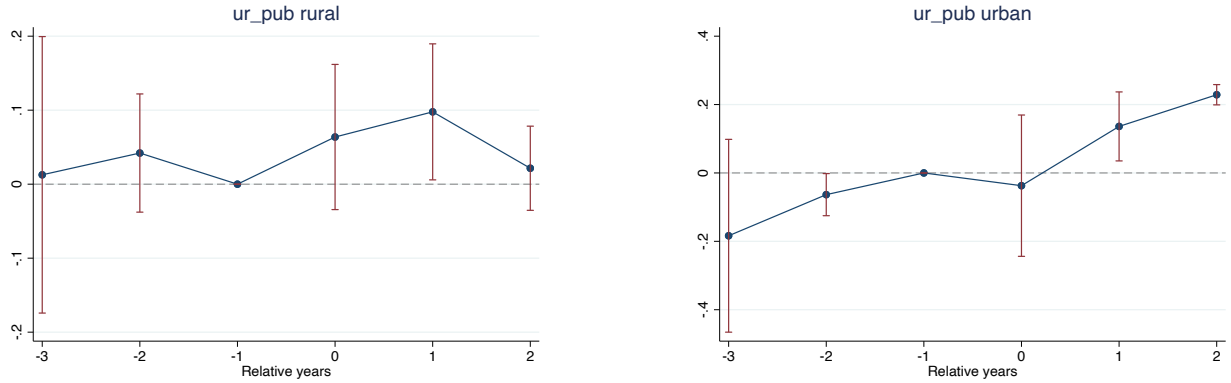


(a) Low education

(b) High education

Notes: See notes to Figure 24. Low education = primary or below; high education = secondary or above.

Figure 28: Event Study: Public Emergency Utilisation by Urbanicity (CdH)

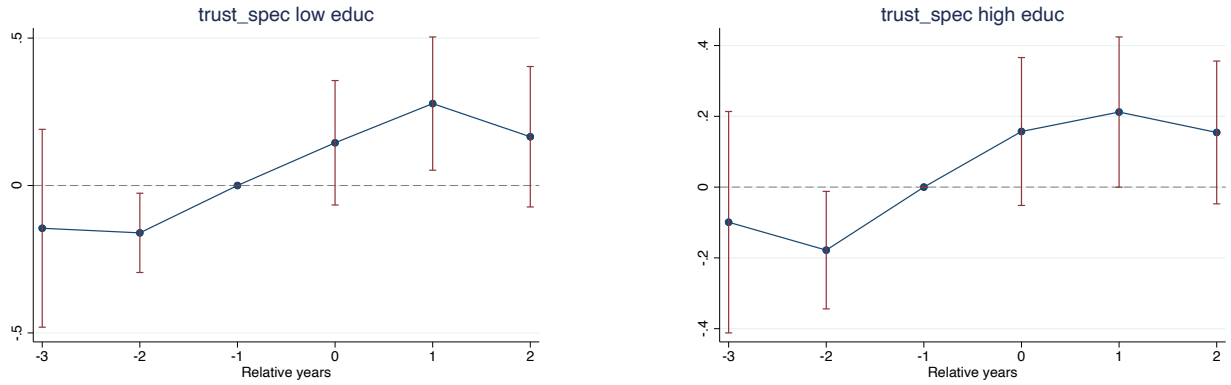


(a) Rural regions

(b) Urban regions

Notes: See notes to Figure 24. Urban regions: Andalucía, Cataluña, Comunitat Valenciana, Madrid.

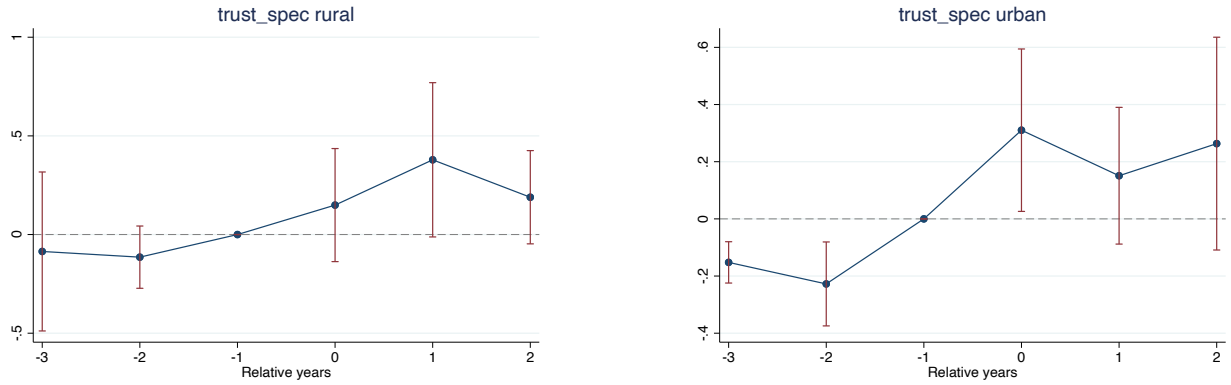
Figure 29: Event Study: Trust in Specialist Care by Education (CdH)



(a) Low education **(b)** High education

Notes: See notes to Figure 24.

Figure 30: Event Study: Trust in Specialist Care by Urbanicity (CdH)



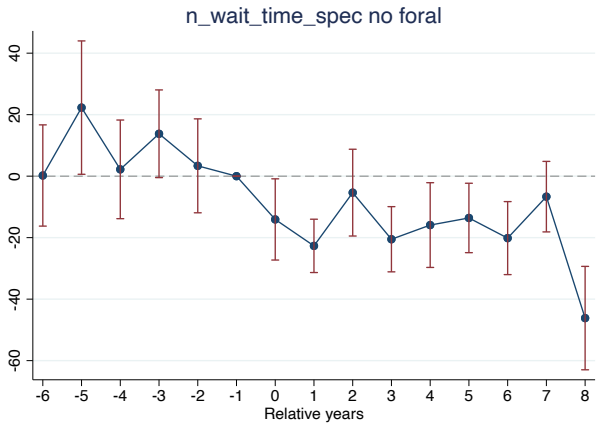
(a) Rural regions **(b)** Urban regions

Notes: See notes to Figure 24.

B.7 Sensitivity: Excluding Foral Regions

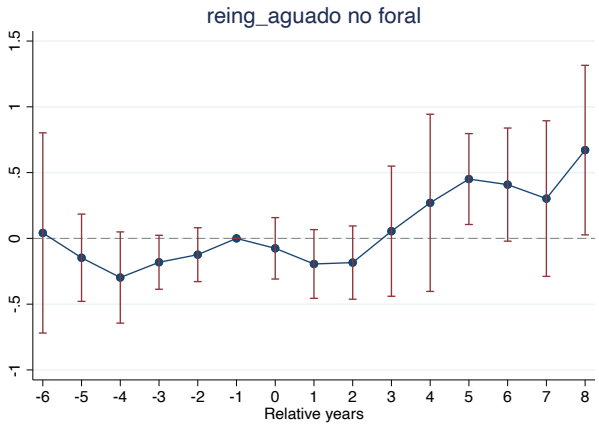
Figures 31–33 report event-study estimates from the specification excluding the two foral regions (Navarra, País Vasco), which have distinct fiscal and health governance arrangements.

Figure 31: Event Study: Waiting Times, Excluding Foral Regions (CdH)



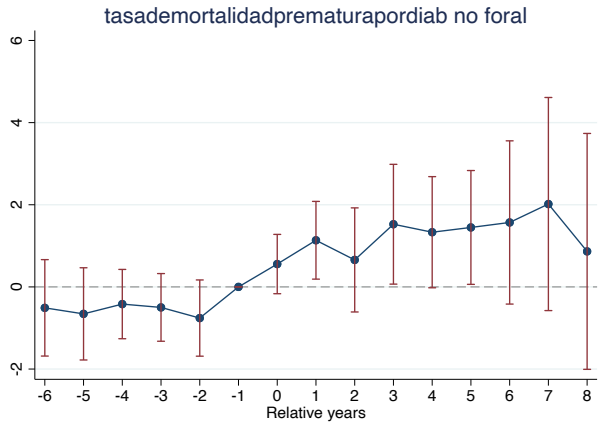
Notes: CdH dynamic effects excluding Navarra and País Vasco. The waiting time reduction is robust to this exclusion.

Figure 32: Event Study: Readmissions, Excluding Foral Regions (CdH)



Notes: CdH dynamic effects excluding foral regions.

Figure 33: Event Study: Diabetes Mortality, Excluding Foral Regions (CdH)



Notes: CdH dynamic effects excluding foral regions. Note the pre-treatment instability, consistent with the mean-reversion interpretation discussed in Section 6.2.1.

C Three-Estimator Comparison

Table 13 presents the aggregated ATT from all three estimators for the main outcomes, with SA standard errors now computed via region-clustered cluster bootstrap in R (Appendix H). The key results are robust: waiting times decline under CdH (-16.98 days) and SA (-8.21 days), and the SA 95% confidence interval contains the CdH point estimate. Private specialist visits decline significantly under both CdH (-0.298) and CS (-0.265). All clinical outcomes remain insignificant across all methods.

Where the estimators disagree—notably for waiting times (CS positive) and public specialist visits (CS negative vs. CdH positive)—the discrepancies reflect known features of the CS aggregation in settings with few treatment cohorts and always-treated units. In all such cases, the event-study representations provide a more informative picture of the dynamic treatment effect than the scalar ATT.

D SDID: Secondary Outcomes and Robustness

This appendix reports supplementary SDID results for the Madrid case study (Section 6.5).

Table 13: Summary of ATT Estimates: Three Robust DiD Estimators

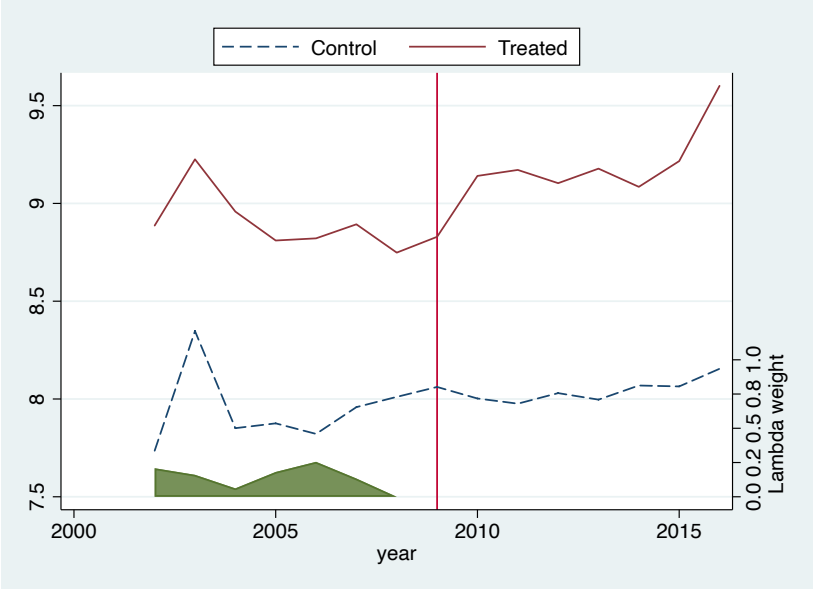
	CdH		CS		SA	
Outcome	ATT	(SE)	ATT	(SE)	ATT	(SE)†
Process						
Waiting time (days)	−16.98***	(2.50)	11.17	(7.81)	−8.21	(7.76)
Satisfaction (0–10)	−0.010	(0.091)	0.080	(0.072)	0.009	—
Clinical						
Readmissions, acute (%)	0.074	(0.175)	−0.201***	(0.069)	0.243	(0.312)
Mort., diabetes	1.114	(0.687)	—	—	1.045	—
Mort., cardiovascular	−0.688	(1.029)	—	—	−0.936	(1.743)
Mort., respiratory	0.939	(0.922)	—	—	0.064	(1.313)
Mort., COPD	−0.076	(0.384)	—	—	−0.422	(0.539)
Infant mortality	0.147	(0.279)	—	—	0.213	(0.315)
Utilisation						
Public specialist visits	0.250**	(0.110)	−0.403	(0.395)	0.032	(0.186)
Private specialist visits	−0.298***	(0.058)	−0.265***	(0.068)	−0.021	(0.082)
Spending						
Health exp. per capita (€)	−0.80	(25.85)	—	—	5.48	—
% exp. on salaries	−0.04	(0.22)	−0.594***	(0.147)	−0.04	—
Day hospital beds	−2.60	(2.61)	−0.378	(6.999)	−1.40	—

Notes: Baseline specification excluding foral regions (Navarra, País Vasco). CdH = de Chaisemartin and D'Haultfœuille (2020); CS = Callaway and Sant'Anna (2021); SA = Sun and Abraham (2021). CdH ATT is the average of dynamic effects; CS ATT from `estat simple`; SA ATT is the average of post-treatment IW coefficients. CdH and CS standard errors (in parentheses) are from wild cluster bootstrap (999 replications) at the region level. † SA standard errors are computed via region-clustered cluster bootstrap with 999 replications implemented in R using `fixest::sunab()` (Appendix H); the SA ATTs in this column are the bootstrap-averaged point estimates, marginally different from the single-shot Stata estimates. "—" indicates that the estimator failed to converge for that outcome or the SE is unavailable in the bootstrap (typically because of insufficient post-treatment overlap). Spending outcomes are estimated without public health expenditure as a control (since it is itself a spending variable). Results including foral regions are virtually identical and available upon request. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

D.1 Secondary Outcomes

Figure 34 presents the SDID trajectory for public specialist visits. Consistent with the main CdH estimates in Table 3, the SDID suggests a positive effect on public specialist utilisation following Madrid's 2009 reform, mirroring the decline in private visits.

Figure 34: SDID Trajectory: Madrid Public Specialist Visits

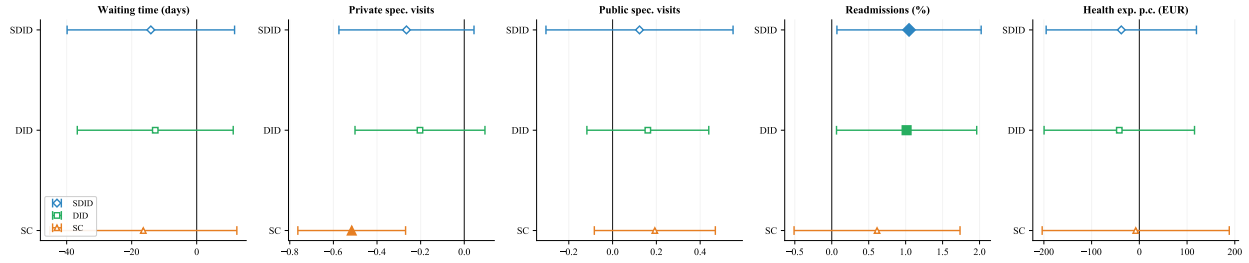


Notes: See notes to Figure 9.

D.2 SDID vs. DID vs. SC Comparison

Figure 35 compares the SDID, standard DID, and synthetic control estimates for Madrid across all five outcomes. The three methods produce broadly consistent estimates for waiting times and private specialist visits—the two primary outcomes. Differences across methods are larger for secondary outcomes (readmissions, health expenditure), where the point estimates are small and insignificant under all three approaches.

Figure 35: Madrid Case Study: Comparison of SDID, DID, and SC Estimates



Notes: SDID = Synthetic DiD (Arkhangelsky et al., 2021); DID = standard difference-in-differences; SC = synthetic control. All estimates for Madrid (treated 2009) vs. never-treated donor pool. Horizontal bars: 95% CIs from placebo inference (50 replications). Filled markers indicate significance at 5%.

D.3 Restricted Donor Pool

As a robustness check, we re-estimate the SDID for Madrid's two primary outcomes excluding later-treated regions (Valencia, Galicia, La Rioja) from the donor pool, retaining only the seven never-treated regions. This ensures that the synthetic counterfactual is constructed exclusively from units that never adopted provider choice reforms. The restricted-donor SDID estimates remain negative for both waiting times and private specialist visits, confirming the robustness of the Madrid case-study findings. Full results are reported in the log file of `07_madrid_sdid.do`.

E HonestDiD Sensitivity Analysis

We implement the Rambachan and Roth (2023) sensitivity analysis for the specialist waiting time result. The HonestDiD framework asks: how large would violations of parallel trends need to be to overturn the conclusion that the ATT is negative?

The smoothness restriction bounds the change in the pre-trend violation between consecutive periods by a parameter $\bar{M}$. At $\bar{M} = 0$ (exact parallel trends), the robust confidence interval for the waiting time ATT is $[-22.1, -11.9]$, excluding zero. As $\bar{M}$ increases, the confidence interval widens. The main result survives up to $\bar{M} = 2.0$, meaning that

even if the pre-trend violation accelerated by up to 2 days per period, the estimated effect remains significantly negative. Given that the observed pre-treatment coefficients show no systematic trend (joint F-test: $p = 0.163$), this threshold is conservative.

For private specialist visits, the robust confidence interval excludes zero up to $\bar{M} = 1.5$, providing additional assurance for the second key finding.

F Data Coverage

Table 14 documents the region-year availability of specialist waiting time data, the key outcome with reduced sample size ($N = 221$ vs. $N = 322$).

Table 14: Specialist Waiting Time Data Availability by Region

Region	Treatment status	Years available	Years missing	N
Andalucía	Never treated	2002–2019	—	18
Aragón	Treated (2007)	2003–2019	2002	17
Baleares	Never treated	2004–2019	2002–2003	16
Canarias	Never treated	2002–2019	—	18
Castilla-La Mancha	Never treated	2003–2019	2002	17
Castilla y León	Never treated	2002–2019	—	18
Cataluña	Never treated	2002–2019	—	18
Galicia	Treated (2015)	2002–2019	—	18
Madrid	Treated (2009)	2002–2019	—	18
Valencia	Treated (2015)	2002–2019	—	18
La Rioja	Treated (2016)	2007–2019	2002–2006	13
Asturias	Never treated	2007–2019	2002–2006	13
Cantabria	Never treated	2007–2019	2002–2006	13
Extremadura	Never treated	2006–2019	2002–2005	14
Murcia	Never treated	2010–2019	2002–2009	10
Navarra	Always treated	—	2002–2019	0
País Vasco	Always treated	2005–2019	2002–2004	15
Total region-year observations				221

Notes: Missing data reflect the phased rollout of the INCLASNS reporting system. All five treated regions (excluding foral) have complete data from at least 2003 onward, covering the full pre-treatment and post-treatment windows relevant for estimation.

G Randomization Inference

With only 17 regions, the asymptotic justification for cluster-robust inference is weak. Wild cluster bootstrap with $B = 999$ replications addresses one side of the problem (Cameron et al., 2008; Webb, 2023), but a complementary check is to assess whether the observed effects could arise by chance under a sharp null of no treatment effect for any region. Following Young (2019), we implement Fisher randomization inference (RI).

Procedure. For each main outcome, we (i) compute the observed ATT under the headline CdH specification (`did_multiplegt_old` with `dynamic(5)` and region-level clustering); (ii) randomly permute the vector of adoption years across regions $B = 999$ times, preserving the empirical distribution of treatment timing (the same five adoption years – 2007, 2009, 2015, 2015, 2016 – redistributed across regions); (iii) re-estimate the CdH ATT under each permutation; and (iv) report the two-sided RI p-value as the share of permuted ATTs at least as large in absolute value as the observed one. Foral always-treated regions (Navarra, País Vasco) are excluded throughout, since they have no pre-treatment period in our sample window and so cannot be assigned a counterfactual adoption year. Replication code is in `stata_RR/08_RR_randomization_inference.do`.

Results. Table 15 reports observed CdH ATTs, wild cluster bootstrap p-values, and RI p-values for the three regional outcomes that move the substantive interpretation. The two pre-registered effects – waiting times and private specialist visits – are confirmed by RI at conventional thresholds: $p_{\text{RI}} = 0.025$ for waiting times and $p_{\text{RI}} = 0.008$ for private specialist visits. The estimated increase in public specialist visits, by contrast, is not distinguishable from chance under permutation ($p_{\text{RI}} = 0.263$), which is consistent with our Section 6.4 finding that this effect is essentially Madrid-driven. Figure 36 shows the permutation distributions for waiting times and private specialist visits, with the observed ATT marked. In both cases the observed effect falls in the tail of the empirical null distri-

bution.

Table 15: Randomization inference for the three main regional outcomes

Outcome	ATT (CdH)	p, bootstrap	p_{RI}	Permutations
Specialist waiting time	-15.47	< 0.001	0.025	999
Public specialist visits	$+0.194$	0.023	0.263	999
Private specialist visits	-0.292	< 0.001	0.008	999

Notes: ATT (CdH) is the average post-treatment dynamic effect from `did_multiplegt_old` with `dynamic(5)` and region clustering, on the panel excluding the foral always-treated regions (Navarra, País Vasco). The observed ATT differs slightly from the pooled headline (e.g., -15.47 vs. the pooled -16.98 for waiting times) because the pooled specification includes the foral regions through their always-treated status. The wild cluster bootstrap p-value uses $B = 999$ replications at the region level; the RI p-value is the two-sided share of permuted ATTs at least as large in absolute value as the observed.

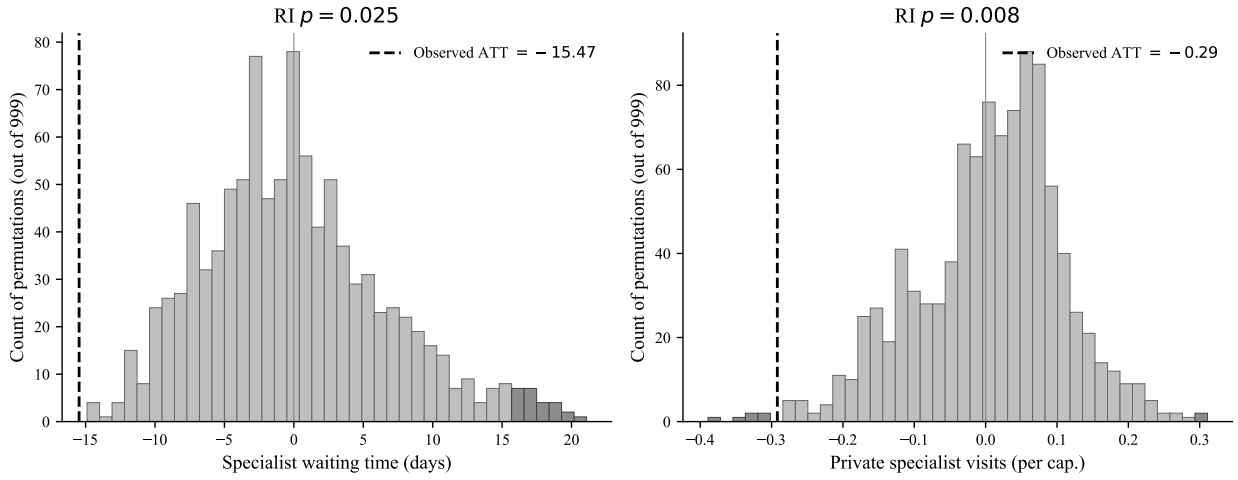


Figure 36: Randomization inference: permutation distribution of the CdH ATT under the sharp null of no treatment effect, $B = 999$ region-level permutations of adoption years. The vertical dashed line marks the observed ATT; bars shaded darker correspond to permutations at least as extreme as the observed effect (which determine the RI p-value).

Interpretation. The wild cluster bootstrap and RI deliver concordant conclusions on the two main effects, neither rejecting public-visit reallocation as an independent effect. This convergence alleviates the concern that conventional inference is mis-calibrated by the small number of clusters: the headline effects on waiting times and private specialist

visits are not artefacts of inference assumptions and survive a permutation test that makes no parametric assumption about the within-region variance structure.

H Alternative Estimator Details

The three estimators address different aspects of the bias arising in staggered DiD settings with heterogeneous treatment effects.

The de Chaisemartin and D'Haultfœuille (2020) estimator (CdH) decomposes the TWFE estimand into properly weighted group-time treatment effects, comparing only "switchers" against stable units. The dynamic specification provides post-treatment effects and pre-treatment placebo coefficients, with inference based on a cluster wild bootstrap. This is our primary estimator because it handles always-treated units transparently and provides interpretable dynamic effects even with few treatment cohorts.

The Callaway and Sant'Anna (2021) estimator (CS) constructs group-time ATTs and aggregates them using only not-yet-treated units as controls. The doubly-robust IPW procedure provides efficiency gains. In our application, the CS estimator produces stable estimates for outcomes with sufficient within-group variation (waiting times, utilisation, satisfaction, readmissions) but fails to converge for several mortality indicators. We attribute this to the small number of treatment cohorts (5) and the limited within-group variation in mortality rates.

The Sun and Abraham (2021) estimator (SA) re-weights TWFE event-study specifications by cohort-specific effects, removing contamination from already-treated controls. In our application, the Stata routine `eventstudyinteract` produces event-study coefficients but does not populate the off-diagonal entries of the variance-covariance matrix for the aggregated post-treatment ATT (`e(V_iw)`), preventing direct delta-method computation of a standard error for the scalar ATT. We obtain a valid SE via region-clustered cluster bootstrap implemented in R using `fixest::sunab()`. The procedure: in each of $B = 999$

replications we draw regions with replacement, re-estimate the SA event study with new region IDs, average the post-treatment event-study coefficients to form the aggregated ATT, and store the value. The empirical SD of the bootstrap distribution is reported as $\widehat{\mathrm{SE}}(\bar{\beta}_{\mathrm{post}})$ and the 2.5–97.5 percentile interval as a placebo-style 95% CI. The SA cluster-bootstrap SEs populate the SA columns of Table 13.

H.1 Why the CS Aggregated ATT Reverses Sign on Waiting Times

A reasonable question is why, mechanically, the CS aggregated ATT yields a positive number for waiting times (+11.17 days, $p = 0.152$) when both CdH and SA, and indeed the CS event-study coefficients themselves, point negative. The answer lies in the cohort weighting.

The CS aggregated ATT is a weighted average of group-time ATTs $\widehat{\mathrm{ATT}}(g, t)$ with cohort g indexing the year of adoption. CS weights each (g, t) contribution by cohort size and exposure period and uses *not-yet-treated* units as the comparison group. Three features of our design interact to produce the observed sign reversal. The always-treated foral regions (Navarra and País Vasco), which had patient choice in place before our sample begins, are dropped by CS from both the treated and the control sets. The not-yet-treated comparison pool then shrinks rapidly as the five adoption cohorts (2007, 2009, 2015, 2015 and 2016) successively activate, so that by the late 2010s the available controls are essentially the ten regions that never adopt. Madrid, finally, is the largest treated region and contributes the longest post-treatment run, so the CS aggregator places most of its weight on the early-cohort/late-period block in which Madrid is the dominant treated unit.

The aggregator therefore tilts toward a comparison in which the late never-adopters serve as controls for a Madrid-dominated treated group during years (2010–2014) when several never-adopters experienced their own slightly upward-trending waiting-time series. The individual $\widehat{\mathrm{ATT}}^{CS}(g, t)$ entries are mostly negative post-treatment (visible in the

event-study representation, Figure 6, red circles), but the slice with the largest weights pulls the aggregated scalar toward zero or above.

CdH avoids this because it forms the aggregated effect by averaging dynamic event-time coefficients with weights that depend on the number of treated regions at each event time, not on cohort size or comparison-pool size. The SA estimator likewise re-weights the event study by cohort-specific shares without inheriting the not-yet-treated comparison shrinkage. The CS sign reversal is therefore a known feature of the aggregator in small-cohort, always-treated, Madrid-dominated panels rather than evidence against the waiting-time effect. We retain the CS column in Table 13 for transparency and read the CS *event study* as the more reliable visualisation alongside the CdH headline scalar.

I Per-Outcome Minimum Detectable Effects

To assess whether the null clinical results are economically informative or merely reflect insufficient statistical power, we compute minimum detectable effects (MDEs) for each clinical outcome at 80% power and a 5% two-sided significance level. We follow the standard cluster-level DiD formula (Bloom, 2007):

$$\text{MDE} = (z_{1-\alpha/2} + z_{\gamma}) \sqrt{\frac{2}{N\,\pi\,(1-\pi)}}\,\sigma_e,$$

with $(z_{1-\alpha/2}, z_{\gamma}) = (1.96, 0.84)$ the standard-normal critical values for 5% significance and 80% power; $N = 17$ the number of clusters (regions); $\pi = 0.108$ the share of treated region-year cells in the post-reform panel; and σ_e the standard deviation of the outcome *after* partialling out region and year fixed effects. The formula gives the MDE for a static treated $\times$ post DiD coefficient and therefore acts as an upper bound on the power of any heterogeneity-robust event-study estimator built on the same within-region post-FE variation, including the CdH-aggregated ATT we report (Black et al., 2022).

Why the residual SD, not the between-region SD. In an event-study DiD the identifying variation is the within-region post-FE residual, not the cross-sectional dispersion across regions. Using the between-region SD overstates power, because it includes time-invariant region differences that the fixed effects absorb, and yields artificially small MDEs. We therefore use the partialled-out residual SD, computed from a regression of each outcome on region and year dummies on the full pre-reform panel. The originally submitted draft used the cross-sectional SD; this produced MDEs roughly half as large (for instance, 3–4 per 100,000 for cardiovascular mortality rather than the corrected 8.36).

Classification rule. We classify a null as *informative* when its MDE is of the same order of magnitude as the effect documented for the closest comparable reform in the literature — concretely, when MDE is at most twice the relevant comparator effect. We classify it as *power-constrained* otherwise. The threshold is qualitative but applied uniformly; we err on the conservative side. The most informative comparators for our setting are the English NHS choice-and-competition reforms studied by Cooper et al. (2011) and Gaynor et al. (2013).

Table 16: Per-outcome MDEs for clinical outcomes (80% power, $\alpha = 0.05$)

Outcome	Pre-mean	$\widehat{\sigma}_e$	MDE	MDE / mean
Public satisfaction (0–10)	6.56	0.20	0.57	8.6%
Acute care readmissions (%)	7.42	0.37	1.15	15.5%
Urgent post-surgery readmissions (%)	10.39	1.34	4.16	40.1%
Urgent surgical/transplant readm. (%)	2.65	0.17	0.52	19.5%
Cardiovascular mortality (per 100k)	28.10	2.70	8.36	29.8%
Respiratory cancer mortality (per 100k)	14.98	1.56	4.83	32.3%
COPD mortality (per 100k)	9.11	1.14	3.53	38.8%
Infant mortality (per 1,000)	3.17	0.63	1.95	61.6%

Notes: MDEs computed using the formula in the text with $N = 17$ and $\pi = 0.108$. $\widehat{\sigma}_e$ is the residual SD from regressing each outcome on region and year fixed effects on the pre-reform panel. Under the classification rule (text), *all* clinical outcomes shown are power-constrained: their MDEs exceed the order of magnitude of effects documented in comparable choice/competition reforms.

Interpretation by outcome. Cardiovascular mortality has the smallest MDE-to-mean ratio in the table (29.8%) and the strongest empirical comparator. Cooper et al. (2011) estimate a 0.31 percentage-point per-year reduction in 30-day in-hospital AMI case fatality per standard deviation of competition in the English NHS; applied to typical UK AMI incidence rates, this implies a population-level mortality reduction of approximately 0.6–0.8 per 100,000 — an order of magnitude smaller than our 8.36 per 100,000 MDE. Even the most powerful clinical outcome in our design cannot rule out effects of the magnitude found in the English reforms. Readmission outcomes are more power-constrained still: detecting a true effect would require shifts of 1–4 percentage points, larger than any documented in comparable choice reforms. Public satisfaction has the smallest absolute MDE (0.57 points) but is more than an order of magnitude above the point estimate we observe (-0.010).

Implications for the theoretical contribution. The incomplete-contracts framework (Holmström and Milgrom, 1991) predicts that demand-side pressure without activity-based funding should improve contractible process quality while leaving non-contractible clinical outcomes unaffected. A confirmation of this prediction requires *precisely estimated zeros* on the clinical side — not merely non-significant point estimates from a low-powered design. Our null clinical results are therefore best read as *consistent with* the prediction rather than confirmation of it. Only the waiting-time and demand-reallocation findings have MDEs tight enough relative to plausible effect sizes to support strong claims; the clinical estimates contribute to the paper's argument only by failing to overturn the framework, not by validating it. Detecting clinically meaningful effects in this setting would require substantially more clusters, within-region patient-level data, or longer post-treatment horizons.