Looks Expensive Until It Doesn't: The Timing of Pain-and-Palliative Care in Colombia

Juan P. Aparicio[1] Diana Echavarria[2]

Abstract

Health systems are expanding access to pain-and-palliative specialist care partly on the promise that earlier contact lowers recorded service use. Administrative data make that promise hard to test: patients are referred precisely when their health, and their recorded spending, is escalating, so the answer depends on what "earlier" is compared against. Using Colombia's national health claims (RIPS, 2015-2021), we show how much it depends. Compared with their own past, the 136,113 patients we study look 28.6 percent more expensive after their first specialist consultation. Matched instead to similar patients whose first visit is a median of 19 months away, the sign reverses: earlier access is associated with about a quarter lower recorded consultation and procedure values per patient-month, with much of the decline concentrated in a high-spending tail. We then convert the estimate into a benchmarked review list. Trained on each patient's matched gap, a machine-learning model ranks later cohorts from pre-referral histories; its flagged top decile shows reductions nine times the cohort average, though that tail has thinner matched support. A prior-spending rule captures most of this value, so the lesson is not black-box personalization but a one-line rule for deciding which histories warrant clinical review first.

Keywords: pain-and-palliative care; treatment timing; risk-set matching; placebo tests; machine learning

JEL Codes: I11; I13; C21; C23; C53; O54

[1]Corresponding author. Institute for Replication and University of Ottawa. Email: juposada93@gmail.com.
[2]Clinica del Rosario.

1. Introduction

Consider a patient in their late fifties with worsening back pain, whose imaging, consultations, and medication adjustments have accumulated for months before anyone refers them to a pain-and-palliative specialist. The referral arrives at the peak of an escalation: compare their spending after the visit with their spending before it, and the specialist looks expensive—not because the visit caused anything, but because the months before it were quiet relative to the crisis that triggered it. This is not one patient's quirk. In Colombia's national claims between 2015 and 2021, the 136,113 patients we study look 28.6 percent more expensive after their first pain-and-palliative consultation than before; compared instead with matched counterparts whose own referral is still more than a year away, they show about a quarter lower recorded service values. The same event, against two defensible-sounding baselines, yields opposite answers—and health systems deciding whether to expand earlier access are reading evidence built on both.

The stakes are large. The need for palliative care is growing fastest in low- and middle-income countries (Knaul et al., 2018; Sleeman et al., 2019), access is a legal entitlement in Colombia (Law 1733 of 2014), and the cost claims circulating in policy debates rest almost entirely on observational comparisons in which referral timing tracks clinical deterioration (May et al., 2018; Morrison et al., 2008; Seow et al., 2022)—the same confound familiar from the health-economics literature on ageing and end-of-life spending, where proximity to death, not time itself, drives the bulk of costs (de Meijer, Koopmanschap, Bago d'Uva, & van Doorslaer, 2011; Werblow, Felder, & Zweifel, 2007). This paper asks: when observably similar patients reach the same specialist service months apart, is earlier access associated with lower or higher recorded service use, and can administrative data answer the question credibly at all?

Our answer comes from a matched timing design whose diagnostics we test, not merely report. For each calendar month of 2016-2021 we match patients whose first specialist pain-and-palliative consultation occurs that month (code `890243` in CUPS, Colombia's classification of billable services) to similar patients in the same month whose own first visit is still at least ten months away (a median of nineteen), using only what the records show before the index month: prior spending, hospital days, diagnoses, and acute-care use. On that margin, earlier access is associated with about a quarter lower recorded consultation and procedure values per patient-month, roughly US$13, or 27 percent of the matched-control mean. The estimate keeps its sign and significance across the main robustness restrictions, though the decline concentrates among the highest-spending patients, where matched support is thinnest and we read the evidence most cautiously. Because a matched estimate is only as credible as its diagnostics, we then test the diagnostics themselves, re-running the entire pipeline in shifted worlds where the true effect is zero: there, spurious estimates nearly the size of the headline arise only alongside failing month-by-month pre-period checks, while the real analysis passes them. Finally, a machine-learning layer converts the estimates into a benchmarked review list held to a one-line rule ranking patients by prior-year spending, which nearly matches it.

Three literatures set up the contribution. Randomized trials establish that earlier palliative care improves symptom control and quality of life in advanced cancer (Bakitas et al., 2015; Kavalieratos et al., 2016; Temel et al., 2010; Zimmermann et al., 2014), but they are small, cancer-specific, high-income, and silent on utilization at national scale. The cost literature reports savings from palliative consultation (Brumley et al., 2007; May et al., 2015; May et al., 2018; Morrison et al., 2008; Seow et al., 2022; Sheridan et al., 2021), but it is concentrated

in high-income hospital and cancer settings and rests on before-after or user-versus-nonuser comparisons that the timing problem can sign-flip. The methods literature supplies the tools: risk-set matching for time-dependent treatments (Hade, Nattino, Frey, & Lu, 2020; Li, Propert, & Rosenbaum, 2001; Lu, 2005) and the modern staggered-timing literature (Callaway & Sant'Anna, 2021; Goodman-Bacon, 2021). But applied matched studies typically report diagnostics whose power is never demonstrated, a gap that parallels recent cautions about pre-trends testing in difference-in-differences (Rambachan & Roth, 2023; Roth, 2022). The standard averaged pre-period "placebo" equals zero by construction even in a deliberately corrupted analysis.

Against those antecedents the paper contributes three things. First, it brings national administrative evidence on recorded service values among patients who eventually enter the same specialist service, with a full service-code decomposition of where the decline comes from, to a cost literature with little comparable national evidence. Second, it moves from applying the risk-set design to validating it, with full-pipeline placebos that demonstrate the diagnostics have power, negative controls, sensitivity bounds, survivorship restrictions, and patient-level cluster-bootstrap inference. Third, it treats targeting as a prediction policy problem (Kleinberg, Ludwig, Mullainathan, & Obermeyer, 2015) and then does what deployment proposals rarely do: it benchmarks the algorithm against a one-variable rule, which captures most of its value. We argue this benchmarking should precede any use of prediction in referral workflows.

The estimates are observational, and one assumption separates them from causal ones: that conditional on matched recorded histories, earlier-access patients would have followed the same path as their delayed counterparts had their own visit been delayed. Under it the matched contrast is the average effect of earlier access on the treated. The assumption cannot be verified, because referral responds to clinical information the records do not contain, but it can be probed, and we probe it from three directions—its testable implications, falsification outcomes that selection should move but a pain consult should not (dental and optometry care; Lipsitch, Tchetgen Tchetgen, & Cohen, 2010), and a sensitivity analysis pricing how much hidden selection it would take to overturn the answer. We present the findings as a tested timing comparison and let readers weigh the assumption with that evidence in hand.

2. Institutional setting and measurement

Colombia's insurance architecture creates exactly the timing variation this paper needs. The Sistema General de Seguridad Social en Salud, established by Ley 100 de 1993, is a regulated national health-insurance system rather than a single public provider (Guerrero, Gallego, Becerril-Montekio, & Vásquez, 2011): people enroll with Entidades Promotoras de Salud (EPS), insurer-like plan administrators that organize networks, authorize services, and contract with providers, through a contributory regime for formal-sector workers and a subsidized regime for everyone else. Coverage is high, but timely specialized care depends on geography, provider availability, referral networks, authorization rules, and differences across EPS and departments (Colombia's first-level administrative regions). These frictions make the earlier-versus-delayed margin meaningful: two similar patients can reach the same specialist service many months apart for reasons of access rather than need.

Palliative care occupies a distinctive place in this system. Law 1733 of 2014 recognized access to palliative care for patients with terminal, chronic, degenerative, and irreversible illnesses and

created obligations for insurers and providers. In Colombia this is a single specialty, *dolor y cuidados paliativos*, spanning both chronic non-malignant pain management and end-of-life care, so the patients who reach it are mostly people living with persistent pain, alongside others facing serious or terminal illness. This gap between the statute and the caseload matters for how the paper should be read. Although Law 1733 frames the entitlement around terminal and degenerative illness, the patients who actually carry the specialist first-consult code are predominantly people with chronic pain and musculoskeletal conditions (85.7 percent of our matched sample), with terminal and cancer cases a smaller minority. The evidence below is therefore cleanest, and is best read, as evidence on the timing of chronic-pain specialist referral; the cancer and seriously-ill subgroups, where the end-of-life spending dynamics that dominate the health-economics literature on dying are most relevant, are the part we read descriptively rather than causally. The clinical reason timing might bear on spending is substitution: the specialist consolidates care scattered across imaging, procedures, and several other specialists, managing symptoms directly and displacing the high-cost diagnostic and curative work-up fragmented care accumulates; whether that lowers recorded values is the empirical question. Colombian and regional studies document incomplete and geographically uneven access, with constraints related to specialist supply, provider distribution, opioid availability, training, and health-system fragmentation (Knaul et al., 2018; Pastrana & De Lima, 2022; Pastrana et al., 2020; Sanchez-Cardenas et al., 2024; Vargas-Escobar et al., 2022). Most serious-illness care is delivered outside specialist palliative services, which is why the timing of first specialist contact is itself a policy margin (Kelley & Morrison, 2015; Quill & Abernethy, 2013; Sanders et al., 2024).

The analysis uses RIPS, the Registro Individual de Prestación de Servicios de Salud (Ministerio de Salud y Protección Social, 2022), an encounter and claims registry rather than a clinical chart: it records coded services and reported administrative values, not symptoms, preferences, treatment goals, final paid amounts, or family burden. In Colombia's CUPS service-code nomenclature, code `890243` records a first consultation with a pain-and-palliative specialist—a visit that assesses symptoms, sets a pain-management plan, and opens an ongoing care relationship—`890343` a follow-up by the same service, and `890443` a separate interconsultation request that reaches a different mix of patients, so the main exposure stays focused on `890243` (`890443` is documented in Appendix Table A8).

The main outcome is the monthly sum of administrative values recorded for consultations and procedures, which we call recorded service values (Appendix Table A10). Three caveats matter. First, the measure leaves out medicines and full inpatient facility accounting, so two large categories of serious-illness cost fall outside it—precisely the facility spending to which the palliative-care cost literature attributes much of its measured saving (May et al., 2018; Morrison et al., 2008); the direction of any total-cost effect therefore requires linked pharmacy, facility, and payment data. Second, the values are tariff-like amounts a provider records against the national fee schedule, closer to a billed charge than a final payment, and under capitated contracting (whose mix varies across providers and plans) they can understate resources used. Third, values are in nominal pesos: matching within calendar month removes common price growth from each comparison, though cumulated totals mix price levels across 2016-2021. For scale, 44,875 COP is about US$13 at the 2016-2021 average rate of roughly 3,450 COP per US$1, the conversion used throughout.

A useful benchmark is Sheridan et al. (2021), which compares costs after palliative-care consultation in SEER-Medicare claims using matched controls and pre-consult cost trajectories. The

present study differs in setting, disease mix (their cohort is advanced cancer; ours is mostly pain and musculoskeletal), treatment definition, timing design, and decomposition; the comparison is methodological, not clinical.

3. Data and design

3.1 Sample, exposure, and outcome

The analysis uses the 2015-2021 RIPS consultation and procedure files. The first recorded `890243` event defines each patient's index month. Index months run from January 2016 through February 2021, so that the six pre-index months and the nine post-index months fall inside the 2015-2021 outcome window. Because RIPS does not carry the code before 2015, the first recorded consult is each patient's first identifiable one.

Table 1 summarizes the sample and the outcome scale. The raw consultation file contains repeated `890243` records per patient (684,529 records across 307,990 patients); the analysis uses each patient's earliest record. To keep a handful of extreme billing values from dominating, monthly patient-level values above the 99.9th percentile of positive months are capped at that percentile. The cap sits at 34.19 million COP and touches 7,553 of 7.55 million positive patient-months. Because the cap affects magnitudes, the findings section reports the estimate under alternative caps.

Four terms carry through the paper. *Eligible treated* patients have an index month leaving room for the six pre- and nine post-index months inside the data window; *matched treated* patients are the subset with at least one acceptable not-yet-treated comparison (the rest are *outside matched support* and excluded); each treated–comparison pairing is a *matched control link*; and because one comparison patient can serve several treated patients, the 319,747 links involve only 50,263 *unique control patients*. The matched-design subsection explains how controls are chosen, and Appendix Table A1 shows how similar the arms look.

Table 1: Sample construction and outcome scale.

Domain	Measure	Value
Exposure scan	CUPS 890243 first-consult records, 2015-2021	684,529
Exposure scan	Unique patients with CUPS 890243	307,990
Matched design	Eligible treated patients	233,832
Matched design	Matched treated patients	136,113
Matched design	Eligible treated patients outside matched support	97,719
Matched design	Matched control links	319,747
Matched design	Unique matched control patients	50,263
Matched treated outcome scale	Mean monthly recorded value, months -6 to -1	96,666
Matched treated outcome scale	Mean recorded value in index month	207,645
Matched control outcome scale	Mean recorded value in index month	211,930
Matched control outcome scale	Mean monthly recorded value, months 1-9	165,294

Notes: Values in nominal COP. Source: authors' calculations from 2015-2021 RIPS consultation and procedure files.

The matched sample is predominantly female (76.6 percent), mean age 58.6, with the most common index diagnoses `R52` pain, `M54` dorsalgia, `M51` disc disorders, `M79` soft-tissue disorders, `M19` arthrosis, and `C50` breast cancer—primarily a pain and musculoskeletal cohort (M-codes musculoskeletal, R52 unspecified pain) with a smaller cancer subgroup. Two departments account for 72.4 percent of matched treated patients (Antioquia, whose capital Medellín is a major specialist hub, 43.7 percent, and Bogotá 28.7 percent), reflecting the geographic concentration of specialist supply; the findings report estimates dropping each major department.

3.2 Why before-after misleads, and what to compare instead

Before any adjustment, the raw data already tell the back-pain patient's story at scale. Among matched treated patients, average monthly recorded values run about 97,000 COP in the six months before the index consult, leap to about 208,000 in the consult month, and settle at about 124,000 over months 1-9, so comparing after with before, the first visit appears to *raise* monthly values by 28.6 percent. The shape (quiet months, a spike at referral, an elevated aftermath) is exactly what referral-at-escalation produces on its own, with no effect of the visit, and echoes the long-standing finding that proximity to death, not age or the calendar, explains much of medical spending (Seshamani & Gray, 2004; Zweifel, Felder, & Meiers, 1999). The rise is the reason a comparison group is needed, not a finding (Figure 1); the findings section sets this same path against matched counterparts, and the sign reverses.

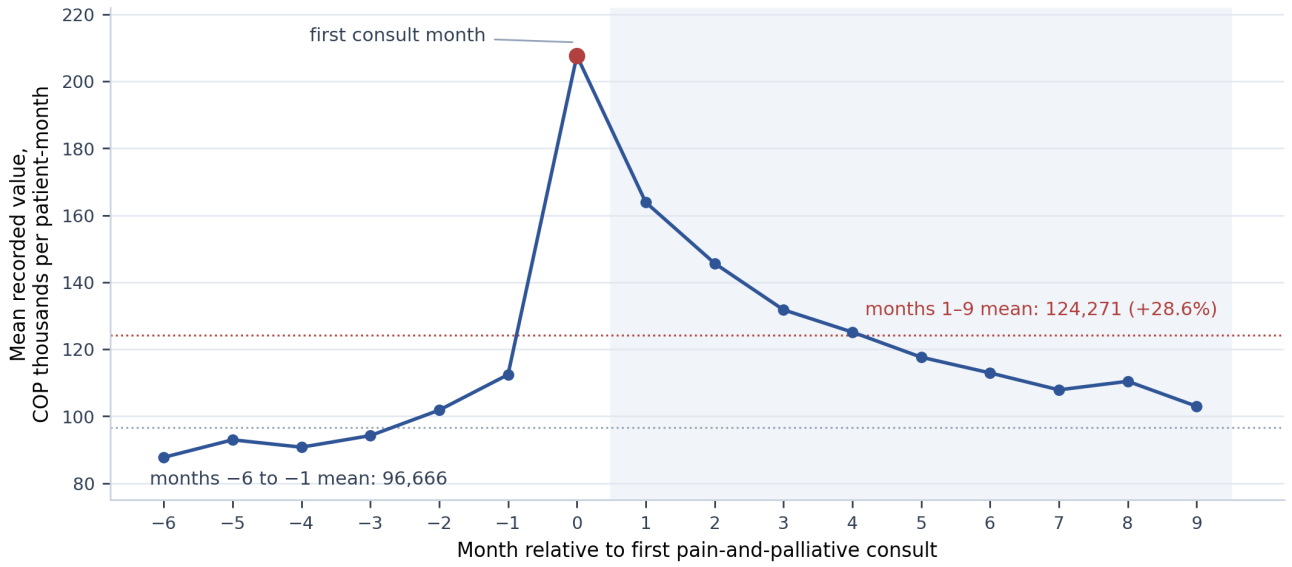

Figure 1: Recorded values of matched treated patients around the first consult. Mean monthly values average 96,666 COP over months -6 to -1, spike to 207,645 in the consult month, and settle at 124,271 over months 1-9 (28.6 percent above the pre-period mean): the path referral at escalation produces on its own.

The baseline we use instead is patients like these whose own referral has not yet arrived. Every control eventually receives the same specialist code, so the comparison's "how much later" is measurable: across matched control links the realized delay between the matched index month and the control's own first consultation has a median of 19 months, a mean of 21.8, and an

interquartile range of 14 to 27, with only about a fifth at the design's 10-12-month minimum and more than half above 18. The paper thus estimates what is associated with seeing the specialist now rather than, for the typical comparison, about a year and a half from now, among patients who all eventually get there and are matchable on recorded histories.

Three boundaries deserve emphasis. First, it covers only patients for whom a comparable delayed control exists: 97,719 eligible treated patients (41.8 percent) had no such match and are excluded, differing from the matched sample by sex, diagnosis, and geography. Second, controls must remain untreated through the follow-up window and be observed again afterward, so the comparison speaks to timing among continuing users; whether that tilts the estimate is tested in the findings. Third, both groups are pinned to a visit—treated patients to the specialist consultation, controls to some other consultation that month—so the comparison never sets patients in active contact against administratively silent ones.

3.3 The matched timing design

The design compares each newly referred patient with the counterpart not referred yet but who will be. Take a patient in their late fifties in Antioquia whose first `890243` consultation occurs in July 2018. The design searches July 2018 for patients of the same sex, age group, department, and three-character diagnosis group who also had a non-specialist consultation that month, whose recent histories match this patient's—deterioration profile, hospital-days trajectory, and recorded spending over the previous six months, and acute-care activity over the most recent three—and whose own first `890243` visit does not arrive until after the nine-month follow-up window closes. Up to three such counterparts become matched controls. The matching agrees exactly where exactness is cheap and meaningful (month, diagnosis group, sex, department, age group, coarse deterioration and hospital-use categories) and picks the closest candidates on the rest. Controls must themselves be on worsening trajectories, so the comparison is escalating-patients-referred-now against escalating-patients-referred-later. Everything the matching sees comes from the six pre-index months, and each set's weights sum to one. One worry deserves an early answer: a decline could reflect patients leaving the data through death or disenrollment rather than cheaper care; but controls are observed again after the window by construction, and imposing the same requirement on treated patients barely moves the estimate, so disappearance does not drive the result.

This is risk-set matching in the sense of Li, Propert, and Rosenbaum (2001): each month's comparison pool is everyone not yet treated, so the comparison never contrasts users with people who would never use the service, only earlier users with later ones. The same month-by-month structure sidesteps the now-well-known pitfalls of regression event studies under staggered timing—every comparison stays inside one calendar month, and no one already treated serves as a control (Borusyak, Jaravel, & Spiess, 2024; Callaway & Sant'Anna, 2021; Goodman-Bacon, 2021; Sun & Abraham, 2021). One pitfall remains, specific to this setting: controls nearing their own referral are already escalating, so late follow-up months can be contaminated by the controls' own ramp-up, which the design tests below measure and strip out.

The two groups look alike where the design demands it (Appendix Table A1): the matched arms agree exactly on sex, age group, and index diagnosis, and closely on the history variables (pre-period recorded values differ by 0.011 standard deviations, procedure counts by 0.038, well inside the 0.10 good-balance threshold; the worst variable is 0.227, everything else below 0.11).

The conditioning does most of its work before the estimator runs: against all candidates in their eligibility cells matched treated patients are 51,587 COP per month *cheaper* in the pre-period, against their chosen matches only +3,852, so the history screens remove 92.5 percent of the selection on spending levels and the residual biases the estimate toward zero.

3.4 Estimation and inference

Estimation proceeds patient by patient. Each treated patient's controls are the nearest candidates by a standardized distance over pre-index spending, emergency, and hospitalization histories (Abadie & Imbens, 2006), hospital-use features weighted most heavily, with weights proportional to inverse distance and summing to one. For each treated patient and each month from six before to nine after the index, we take the gap between her recorded values and the weighted average of her controls', then subtract that set's average gap over the six pre-index months—a matched difference-in-differences in the sense of Heckman, Ichimura, and Todd (1997). This subtraction forces the *average* pre-period gap to zero mechanically, so the averaged pre-period check matched studies customarily report is arithmetic, not evidence; what the data can tell us is whether the gap is flat month by month, the pre-period checks or "leads" we hold to an explicit flatness criterion (no monthly lead above one-fifth of the headline, about 9,000 COP). Each patient's months 1-9 estimate is the average of her nine demeaned gaps, and the headline is the unweighted average across the 136,113 matched treated patients. The index month is reported separately because it mechanically contains the consultation; its estimate is -8,137 COP, so including it would strengthen the contrast.

Getting the uncertainty right requires noticing these comparisons are not independent (Bertrand, Duflo, & Mullainathan, 2004): each control serves about six matched sets, about a third of controls later re-enter as treated patients (Appendix Table A2), and the index months span the COVID-19 pandemic. Our primary standard error is therefore a person-level cluster bootstrap (Cameron & Miller, 2015): we re-draw persons with replacement, each carrying everything they contribute as treated comparison and control link, preserving reuse and dual roles. It gives 3,974 for the main estimate—more than twice the naive independent-observations formula, while calendar-month clustering more than doubles it again; significance is unambiguous under all three (Appendix Figure A2). Component and subgroup estimates carry the naive standard errors and should be read with corresponding caution.

4. Findings

4.1 The matched estimate

The headline contrast reverses the before-after story of Figure 1: the same patients who look 28.6 percent more expensive against their own past run about a quarter *cheaper* against matched counterparts whose first visit is still a median of 19 months away. Over months 1-9, total consultation plus procedure values fall by 44,875 COP per patient-month relative to matched delayed-access controls (Figure 2)—about 27 percent of the matched-control mean, or US$13, and about 404,000 COP (US$115) cumulated over the nine months. A stricter version censoring control months within six of the controls' own first visit gives -38,789, so the headline is not driven by controls entering their own referral ramp. The figure shows the pre-period gaps staying flat

and the decline opening immediately after the consult; those flat gaps are the leads scrutinized below.

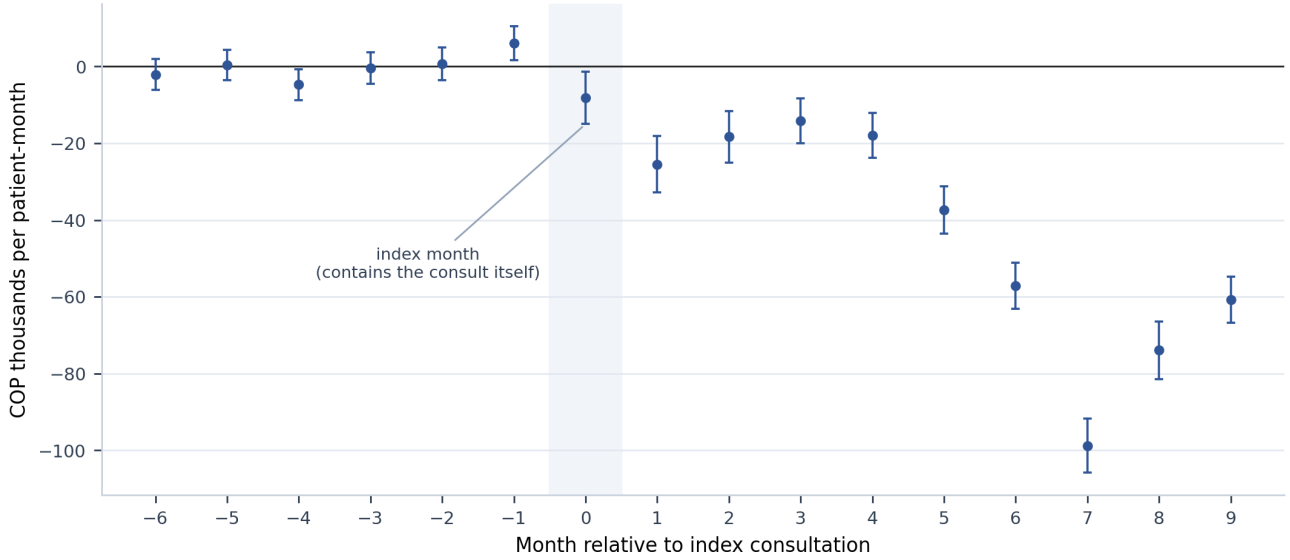

Figure 2: Matched timing estimates by month for total recorded values. Points are matched differences-in-differences net of each matched set's average pre-period difference (months -6 to -1); bars are 95 percent confidence intervals using patient-level standard errors. The shaded month 0 contains the index consultation itself. Takeaway: the pre-period gaps are flat within 6,090 COP, and the post-period decline deepens through month 7 rather than fading.

The decline is not a one-month dip: it starts near 25,000 COP in month 1 and deepens to nearly 99,000 by month 7. Part of the late widening reflects controls' own approach to referral, the contamination the six-month censoring removes (under it the months 1-9 average is -38,789); the remainder is a sustained treated-side decline, with treated monthly values falling monotonically from 163,893 in month 1 to 102,987 in month 9, consistent with consolidation deepening over time rather than a transient lull.

4.2 Where the decline comes from

The single pooled number conceals variation, and the design's checks grade each subgroup. By clinical profile the result is anchored in the pain-and-musculoskeletal core: patients whose index diagnosis falls in the pain or musculoskeletal chapters (85.7 percent of the sample) show -44,170 COP per patient-month with the cleanest pre-period checks in the paper (largest gap 3,747), and patients without any pre-period cancer flag (86.4 percent) show -38,154, also clean (the narrower common pain/MSK groups in Appendix Table A9 cover 80.6 percent). The cancer-related subgroups show larger point estimates (-76,416 for cancer-related index diagnoses, -87,641 for the pre-period cancer flag) but fail the pre-period criterion (largest gaps 79,895 and 43,511), so by the paper's own standard they carry descriptive weight only—and we read them cautiously because, for the most seriously ill, a decline could reflect curative treatment winding down rather than better-coordinated care. The strengthened-observability check bounds this for the sample as a whole: requiring activity at or after month +10 retains 94.9 percent and moves the estimate only to -43,863.

By baseline spending the concentration the prediction section documents is already visible. The top tenth by pre-index recorded values declines about 319,000 COP per patient-month and accounts for roughly seven-tenths of the pooled headline; it is also the group whose pre-period checks are least clean (largest gap 92,160), so we hold it most tentatively. The remaining ninety percent show a smaller decline of about 14,500 with clean leads (largest gap 3,473). The pooled headline thus passes its own checks and its pain-and-musculoskeletal core is solid, so the finding does not rest on the tail, while the very top of the distribution, where matches are thinnest, is best read as suggestive. (Excluding the 12.2 percent of treated patients who also appear as controls leaves the estimate at -44,344, so dual roles do not drive it.) All subgroups appear in Appendix Figure A2.

The decline is as uneven across services as across patients, appearing in both components with procedures accounting for roughly 72 percent: consultation values fall by 12,698 COP per patient-month and procedure values by 32,177. The specialist consult does not reduce its own service line—pain-and-palliative consultation values rise by 4,417, as they should if the visit initiates specialist care—so the net consultation decline comes from consultations outside the specialist service (-17,401), which more than offset the specialist increase (Figure 3; full estimates in Appendix Table A5).

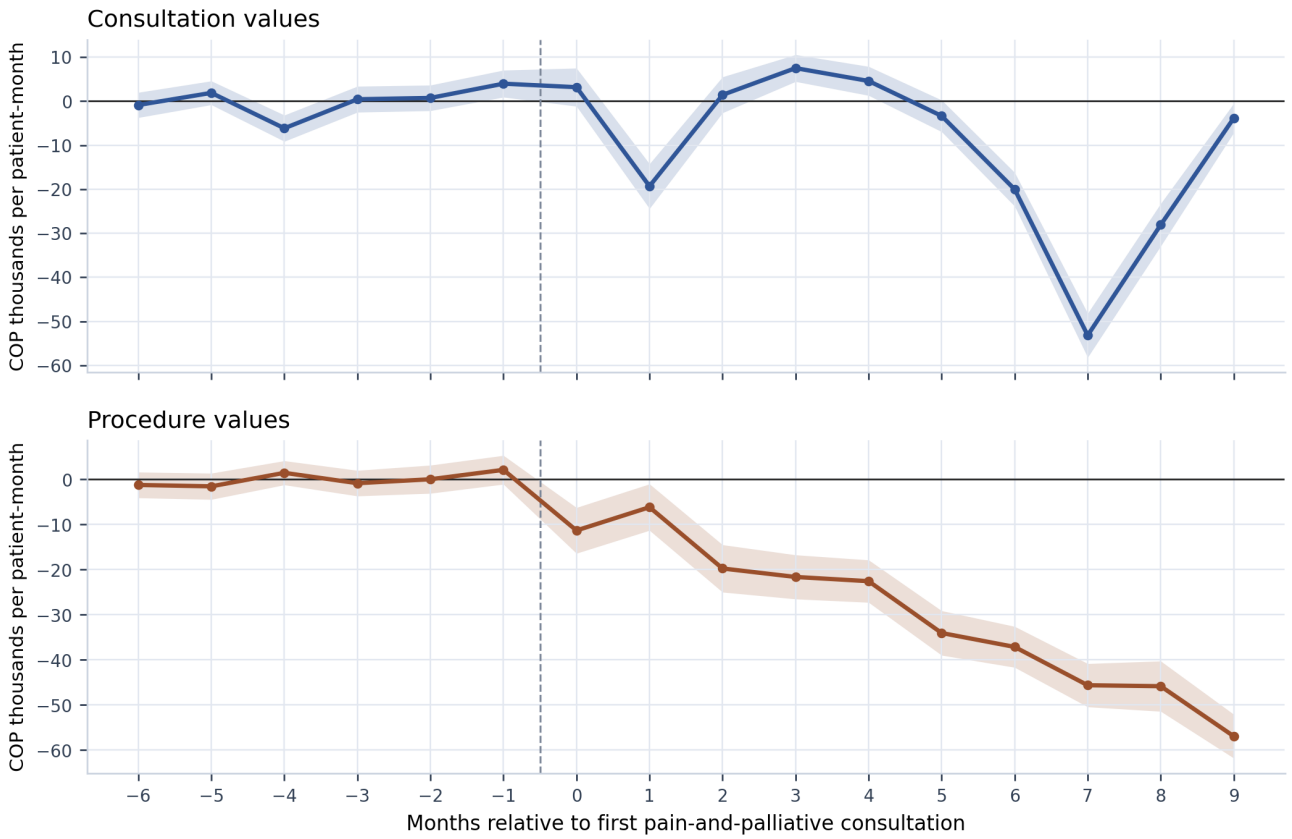

Figure 3: Consultation and procedure recorded-value components. Monthly matched timing estimates for earlier versus delayed first pain-and-palliative consultation, net of each matched set's average pre-period difference; shaded bands are 95 percent confidence intervals.

The code-level decomposition (Appendix Tables A6 and A7) shows which billed services move

(May et al., 2017). Consultation increases concentrate in the specialist service and adjacent care-management contacts; declines, in general-medicine and other-specialty consultations. Procedure increases are pain and symptom-management modalities (nerve injections, regional blocks, neurolysis); declines are high-cost imaging, radiotherapy, antineoplastic therapy, and physical therapy. The pattern is the substitution the mechanism predicts—low-cost targeted pain procedures rising while expensive imaging and curative-cancer therapy fall—though these are billing-code changes, not direct evidence that individual patients appropriately stopped or continued specific therapies.

Part of the decline is missing months rather than cheaper ones, but the smaller part. As the post-period progresses, treated patients grow more likely than their controls to have a month with no recorded activity at all (42.6 versus 34.2 percent by month 9). Decomposing the headline into this extensive margin (fewer billed months) and an intensive margin (lower recorded value per active month) attributes about 22 percent of the pooled decline to fewer billed months and 78 percent to lower value when billing occurs, so consolidation of active care, not disappearance, drives most of the result. The balance flips only in the cancer subgroups, where the extensive margin is the majority (61 percent for cancer-related index diagnoses): disappearance that, in a cohort including seriously ill patients, can reflect unrecorded death or disenrollment—another reason those subgroups carry descriptive weight only, and one the survivorship restriction above bounds for the sample as a whole.

4.3 Testing the design

The credibility of these findings rests on the design, so we examine it directly. A causal reading requires assuming that, conditional on matched recorded histories, earlier-access patients would have followed the same path as their delayed counterparts had their own visit been delayed. This cannot be verified, because referral responds to clinical information the records do not contain, so rather than assert or abandon it we ask three answerable questions: do its observable implications hold, could our diagnostics detect the failures that matter, and how much hidden bias would it take to overturn the answer? Construction details and full tables for every test are in the appendix.

Do the leads hold? The month-by-month pre-period gaps should be flat, by the criterion set with the estimator: no monthly lead above about 9,000 COP. They are (Figure 2): the six range from -4,699 to +6,090 COP, the largest (+6,090 in month -1) being the expected signature of referral at escalation, which if persistent biases the estimate toward zero rather than away. Baselining on months -6 to -3 only, so -2 and -1 become unused checks, leaves the estimate almost unchanged (-43,160; Appendix Table A3).

Could the diagnostics detect a failure? We handed the pipeline worlds where the true effect is zero by construction, shifting every first-consult date earlier so no specialist consultation falls in the outcome window, and re-ran from scratch. A 10-month shift parks the pseudo-window on the pre-referral escalation and manufactures a spurious -42,380 COP, 94 percent of the headline; a 14-month shift, clear of the escalation, returns a statistically-zero +1,329. The monthly leads tell the two apart: the contaminated run fails all six (swinging from -66,819 to +52,679), and no artifact arose with clean leads, while the real analysis (maximum lead 6,090) is the only run with clean leads. The averaged pre-period check, by contrast, is exactly zero in every run including

the corrupted one, which is why we report monthly leads in full rather than the customary one-line summary (Appendix Table A4, Appendix Figure A1). A complementary test confirms the controls' own approaching referrals contaminate late months only modestly: dropping control months within six of a control's own referral moves the estimate to about -39,000, at most one-seventh of the headline.

The battery as a whole holds (Appendix Figure A2): across alternative baselines, visibility restrictions, control-ramp censoring, longer delays, pre-pandemic windows, and dropping the largest department or insurer (one EPS is 41 percent of the sample; dropping it gives -36,822, and every major insurer's estimate is negative), the estimate stays between roughly 32,000 and 45,000 COP of decline, sign and significance never in doubt. The one choice that moves the magnitude is the cap on extreme bills: tighter caps shrink the decline, a looser cap nearly doubles it. Two falsification outcomes with no plausible channel from a pain consult stay quiet: the full estimator on dental consultations returns +21 COP (every pre-period gap inside 39 COP) and on optometry -566, together about one percent of the headline (Appendix Table A15). Selection strong enough to fabricate the result should leak far more.

How much hidden bias? Following Rosenbaum (2002), the largest within-set odds ratio of early referral, Gamma, at which the patient-level declines still reject no effect at 5 percent is 1.48 in the full sample and 1.53 in the pain/MSK core (signed-rank; the appendix reports the sign test). An unobserved factor would thus have to raise the odds of early referral by about half while aligning strongly with subsequent declines. We read this in both directions: the result is not fragile to modest hidden bias, but unrecorded clinical severity, the confounder this design must fear, could plausibly reach that strength, which is why the headline stays a tested timing comparison with the causal reading priced rather than assumed. As a partial probe, linking patients to administrative records back to 2009 (about 94 percent link) shows matched treated and controls balanced on deep pre-2015 hospitalizations and emergency visits (standardized differences -0.002 and 0.05; Appendix Figure A3), outside the window the design matches on; this proxies chronic burden, not the acute deterioration of the referral month, so it narrows but cannot close the assumption.

Taken together, the leads are flat, the placebos show the leads detect constructed failures, two unrelated lines stay quiet, deep histories are balanced, and the estimate survives meaningful hidden-bias stress. None of this turns a matched comparison into a trial; it places the design well inside the range where the timing contrast reads as a credible estimate of earlier access, conditional on the one untestable assumption, rather than as a fragile correlation.

5. From matched estimates to a review list

The matched estimate answers a population question: among comparable patients, is earlier access associated with lower recorded spending on average? A care-management team with capacity to review a fixed number of patients each month faces a different question, which histories warrant review first. This is a prediction policy problem in the sense of Kleinberg et al. (2015): the causal analysis justifies prioritizing clinical review for earlier specialist access, but deciding where to pull that lever is a ranking task rather than a new causal parameter. We therefore use machine learning as a benchmarked review-list exercise, not as proof of individual-level causal effects: the model ranks held-out later cohorts from pre-referral histories, and the

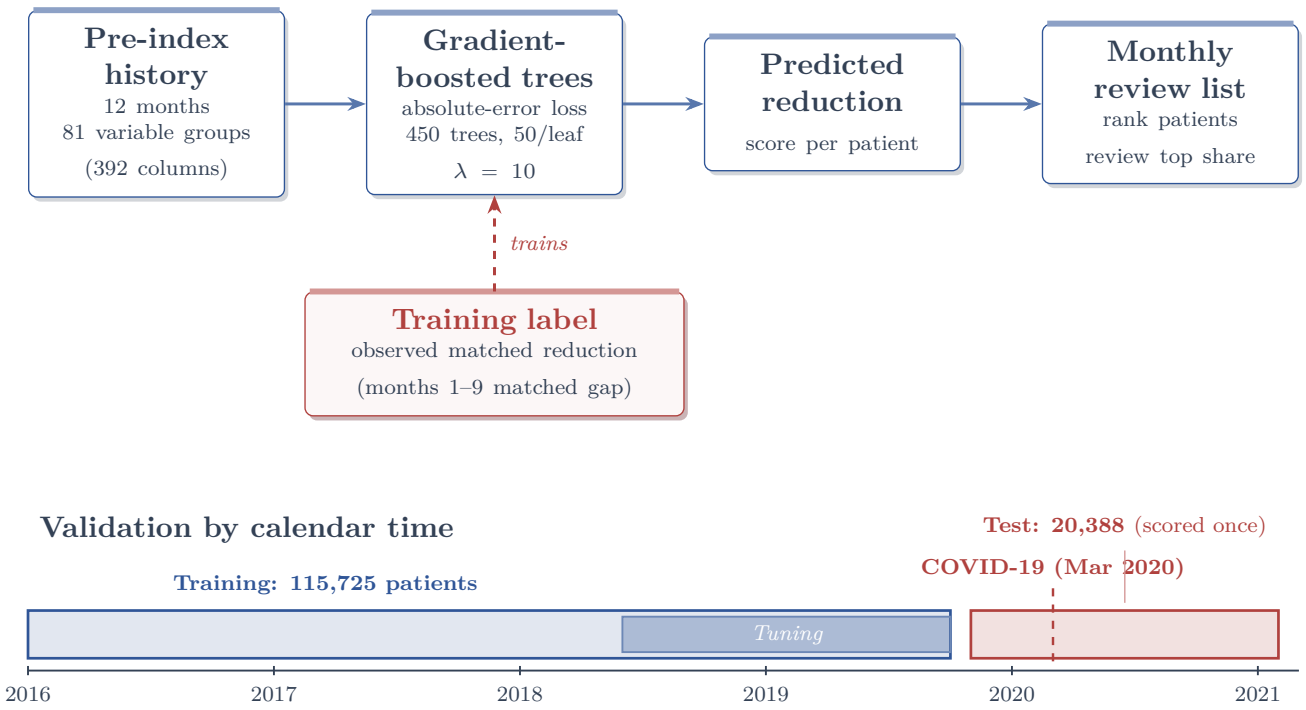

Figure 4: Review-list model and validation design. The model predicts each matched treated patient's observed matched reduction, the months 1–9 matched gap from the causal design (in red), from her pre-index history alone, and the resulting score ranks patients into a monthly review list. Validation is by calendar time: the model trains on January 2016–October 2019 index months, selects its specification on the June 2018–October 2019 tuning cohorts, and is scored once on the November 2019–February 2021 test cohorts, which straddle the pandemic. The training target is winsorized at train-set 1st/99th percentiles.

test is whether that ranking adds anything over a prior-spending rule.

A ranking is only as meaningful as the quantity it ranks on, and the target is where the causal design does double duty. Rather than predict future spending, which would merely re-identify sick patients, the model predicts the object the design already constructed—each patient's months 1-9 matched gap, sign-reversed into an *observed matched reduction* (a recorded-value difference, not money banked). These 136,113 numbers are the training label, and the predictors are confined to the twelve months before the index, so the exercise uses only what would exist before a referral decision.

The modeling choices are conventional. We fit the matched gap with histogram gradient-boosted regression trees (Friedman, 2001), preferred over a linear specification because the signal is interactive—a rising spending slope means something different for a cancer history than for back pain, and geography conditions both (Athey & Imbens, 2019; Mullainathan & Spiess, 2017). Two choices guard against billing-data pathologies: absolute-error loss, so one multi-million-peso month cannot dominate the fit, and regularization (shallow trees, fifty patients per leaf, a learning rate of 0.04 across 450 trees, an L2 penalty) tuned for out-of-sample ranking. The pre-index feature set (81 variable groups, 392 encoded columns) and full design are summarized in Figure 4.

Validation is by calendar time, not a random split, because a deployed list scores later patients:

the model trains on January 2016 to October 2019 index months, selects among twelve candidate specifications on the June 2018 to October 2019 tuning cohorts, and is scored once on the 20,388 patients first seen from November 2019 through February 2021, with even the label-trimming thresholds computed on training rows only. The twelve candidates land their held-out top decile within a narrow band (569,000 to 623,000 COP), so the result does not hinge on the algorithm (Appendix Table A11).

With the specification fixed and the test cohort untouched, the held-out ranking is what the exercise turns on. Sorted by predicted reduction into deciles, the cohort shows the steep gradient a review-first tool needs (Figure 5): the top decile averages about 601,000 COP of observed matched reduction, nine times the cohort mean, with 80 percent positive against a 57 percent base rate, while the gradient falls to zero by the middle deciles and turns negative at the bottom.

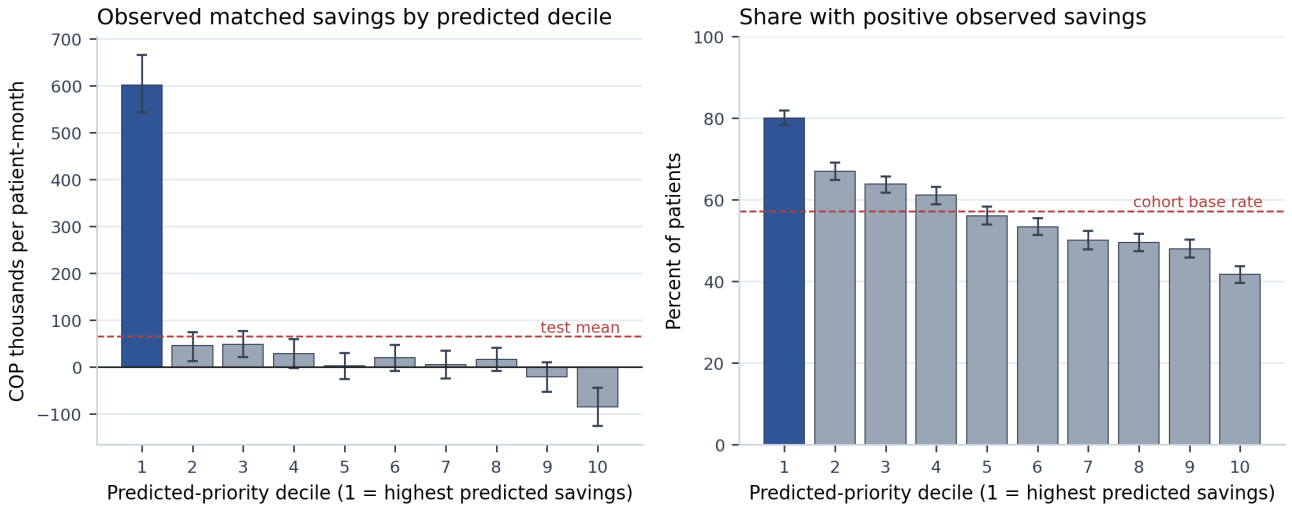

Figure 5: Held-out ranking performance by predicted-priority decile. Left: mean observed matched recorded-value reduction per patient-month in each decile of the held-out test cohorts (November 2019 - February 2021), with 95 percent patient-bootstrap confidence intervals; the dashed line marks the test-cohort mean of 65,510 COP. Right: share of patients with positive observed reductions; the dashed line marks the cohort base rate of 57.1 percent. Positive values mean lower recorded spending than matched controls.

Two facts discipline that gradient. First, it turns the single estimate into a distribution: the -44,875 headline describes almost no one, since reductions concentrate in a high-spending tail, sit near zero through the middle, and turn negative at the bottom, so the policy question is not whether to refer everyone earlier but how to find the small group where the reductions concentrate. Second, the ranking orders groups, not individuals: the model explains only about 14 percent of patient-to-patient variation, what a single noisy realization of a matched contrast yields, so the score ranks groups rather than forecasting any one patient, and any rule must be stated as a share of the list.

The sharpest lesson is the benchmark: the ranking runs on prior recorded spending and essentially nothing else. A rule ordering patients by twelve-month pre-index spending reaches a top-decile mean of about 589,000 COP against the model's 601,000—an edge of roughly two percent (Appendix Table A12) that does not survive moving off the top-decile mean: across the

whole held-out cohort the one-line rule captures at least as much of the total reduction (35.3 versus 34.7 percent in the top decile), concentrates it at least as steeply (concentration coefficient 0.34 versus 0.32), and its rank correlation with realized reductions (0.23) trails the model's (0.27) only marginally; the score correlates 0.85 with that one variable, and a permutation test attributes essentially all of its behavior to baseline spending. A care-management team thus captures nearly all the targeting value with a rule it can compute by hand and audit; the model earns its complexity only at the margin, flagging a lowest decile (about -86,000 COP) for whom earlier access tracks *higher* recorded spending. No individual-level "personalized targeting" claim is demonstrable on claims data of this kind, even as group-level ranking comes nearly free.

That the ranking is, to first order, a baseline-spending ranking is also a caution, because it concentrates flags where mean reversion and matching error are largest. The flagged patients spend about thirteen times the cohort baseline, and their measured reductions equal nearly two-thirds of their own baseline, too large to read as the effect of one consultation. Only 28 percent of the top decile has core (green) matched support, against 43 to 47 percent mid-distribution, and within the top decile the thin-support patients show the larger reductions, so the defensible reading of the top-decile figure is its green-support component, about 436,000 COP (Appendix Table A13). The prediction layer thus hands the same caution back to the causal analysis: the tail it isolates is the lopsided composition already flagged in the findings. In return it shows the ranking is stable out of time (top decile about 507,000 COP before March 2020 and 691,000 during the pandemic, positive in both), though because training outcomes and matched controls straddle the split this is persistence, not independent replication.

Any use in practice is decision support, not an eligibility rule: a flag marks a history resembling those in which earlier access preceded lower recorded values and so merits clinical review, never a reason to withhold referral. Because cost-based targets can embed access bias—patients who use little care because they face barriers look like low-reduction candidates (Obermeyer, Powers, Vogeli, & Mullainathan, 2019)—the flagged group must be audited first; ours over- and under-represents some departments (Appendix Table A14), modest at the top decile but compounding at scale. Any pilot also faces reporting lags, data-protection rules (Ley 1581 de 2012), recalibration as coding drifts (Obermeyer & Emanuel, 2016), and, most consequentially, the supply constraint that makes earlier access for one patient later access for another.

What the layer adds is not a black box standing in for the estimate but a benchmarked, auditable way to turn a population average into a reviewable list—showing where the average comes from and concentrating the expected reductions in a small fraction, out of sample, from pre-referral information alone, while a one-line rule captures most of the value and individual outcomes stay hard to predict.

6. Conclusions

This paper asked whether, when observably similar patients reach the same pain-and-palliative service months apart, earlier access is associated with higher or lower recorded service use, and whether administrative data can answer that credibly at all. The answer depends entirely on the comparison. Against their own past, the 136,113 patients we study look 28.6 percent more expensive after their first visit, because it lands at a moment of escalating need; against matched counterparts whose own referral is still a median of nineteen months away, the sign reverses,

and earlier access is associated with about a quarter less recorded service use per patient-month, concentrated in procedures and non-specialist consultations while specialist contact itself rises. The naive comparison gets the sign wrong; the matched comparison, the one the question intends, gives a bounded answer.

The quarter-less figure becomes the *causal effect* of earlier access under a single assumption: that, conditional on their matched histories, the delayed counterparts show what the treated would have done had their own visit been delayed. Rather than assert it, we test the implications the records reveal—full-pipeline placebos where our monthly lead test catches the constructed failure the averaged pre-period check misses, negative controls that stay flat, and a sensitivity bound showing hidden selection would have to raise the odds of early referral by about half to overturn the sign. None of this proves the assumption, but together it carries the result well past a simple correlation. The same logic governs the review list, where a one-line spending rule does almost as well as the algorithm. The methodological point: for services whose timing tracks deterioration, full monthly leads, full-pipeline placebos, and simple-rule benchmarks belong in the main analysis, not the appendix.

Scope conditions bound the estimate. On external validity, the matched sample covers 58 percent of eligible patients and concentrates in two high-supply departments, so it speaks to well-supplied settings, not the under-served regions where the access question is sharpest. The identifying assumption could still fail through unrecorded clinical severity, which is why we priced it explicitly. The magnitude, though never the sign, depends on how extreme bills are capped. The outcome captures consultations and procedures but not medicines, facility costs, final payments, symptoms, function, or survival, so it is not a total-cost or welfare effect; linked pharmacy, facility, payment, and mortality records are needed to say whether the recorded-value reduction reflects broader savings or better care. Death itself is unrecorded, a first-order concern in cost studies of seriously ill populations given how steeply spending rises near the end of life (Felder, Werblow, & Zweifel, 2010; Wong, van Baal, Boshuizen, & Polder, 2011), though requiring treated patients to be observed again after the window leaves the estimate essentially unchanged. Because recorded values map onto payment differently under capitation, who would capture any savings is unsettled, so we report no aggregate budget figure.

The policy reading is narrow but actionable. Among comparable patients, earlier specialist access is associated with substantially lower recorded service use, and the reduction is large enough, and concentrated enough, to be worth prioritizing for clinical review; whether it also improves welfare turns on outcomes RIPS does not record, such as pain control, function, and survival. One caveat is structural rather than statistical. Because specialist supply is the binding constraint, the estimate is the effect of moving a given patient earlier in the queue, not of lengthening the queue: bringing some patients forward pushes others back, so an earlier-access policy reallocates timing rather than adding capacity, and any system-level saving need not equal the per-patient contrast. Our attempt to find supply-driven timing variation failed (department-month consult throughput predicts almost none of the realized delays), which is one more reason to read the estimate as a timing contrast among comparable patients rather than as the effect of a capacity expansion. The paper's larger contribution is a template for reading administrative data when treatment timing tracks deterioration: choose the comparison group deliberately, then defend it with diagnostics that can fail—full monthly leads, full-pipeline placebos, negative controls, sensitivity bounds, and a benchmarked prediction layer. The natural next steps are linking death and prescription records to widen the outcome and finding settings where specialist supply

shifts for reasons unrelated to patients, so that earlier access can eventually be compared as if assigned at random.

Appendix: robustness and supplementary evidence

This appendix collects the supplementary exhibits, figures first and then tables; both run in the order the text discusses them.

Appendix figures

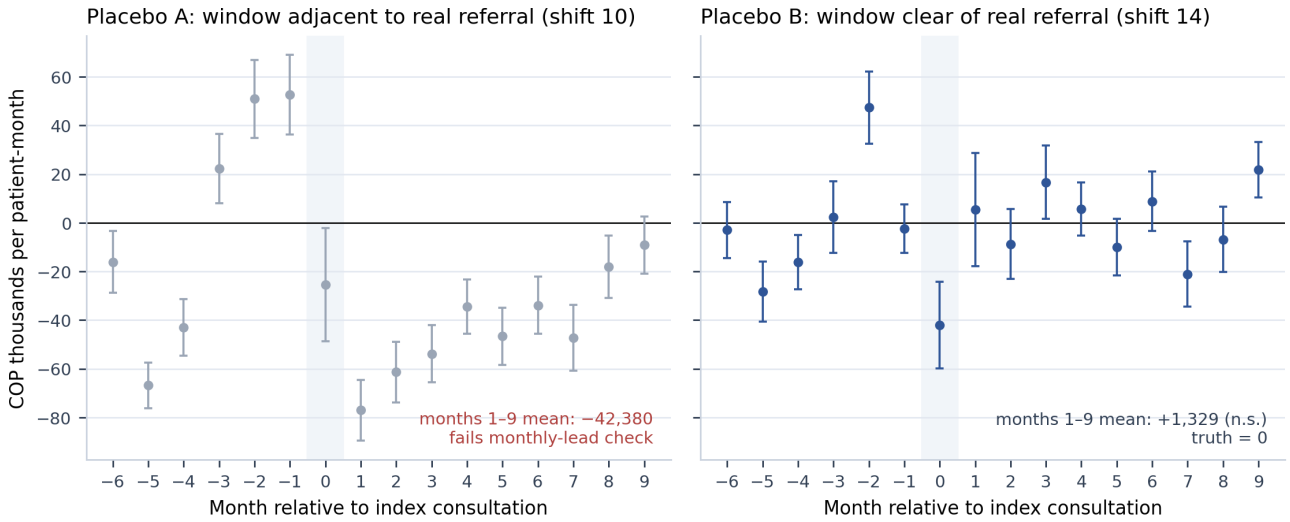

Figure A1: The placebo pair: the same full pipeline in a shifted world where the true effect is zero. Left: with the pseudo-window adjacent to the real referral, baseline contamination produces a large spurious "effect" and wildly unbalanced leads. Right: with the window clear of the real referral, the placebo estimate over months 1-9 is statistically zero. Bars are 95 percent confidence intervals using patient-level standard errors.

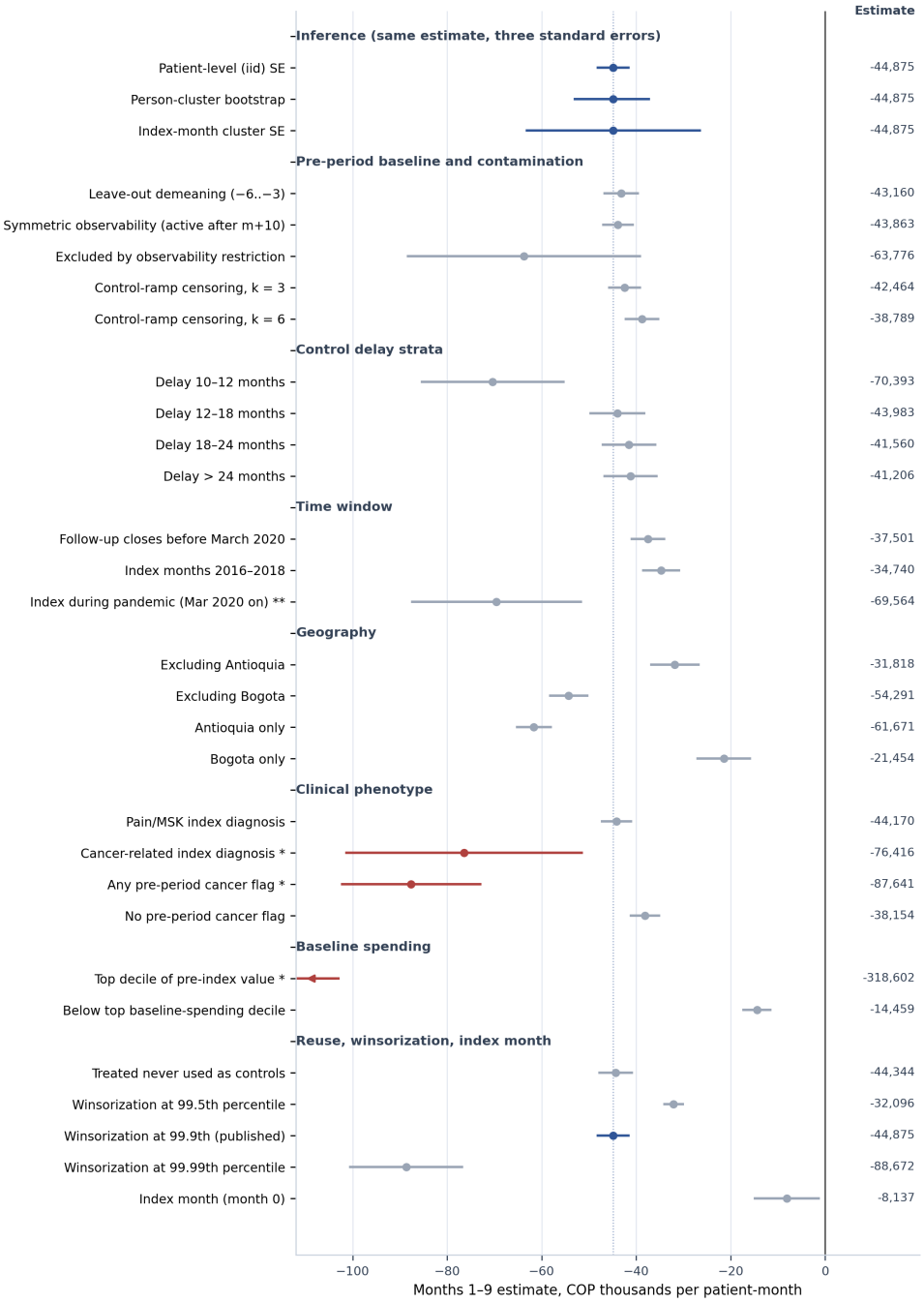

Figure A2: The diagnostic battery: months 1-9 matched timing estimates under alternative specifications. Points are matched differences-in-differences with 95 percent confidence intervals. The first three rows report the headline estimate under alternative standard-error methods; exact estimates are printed at the right.

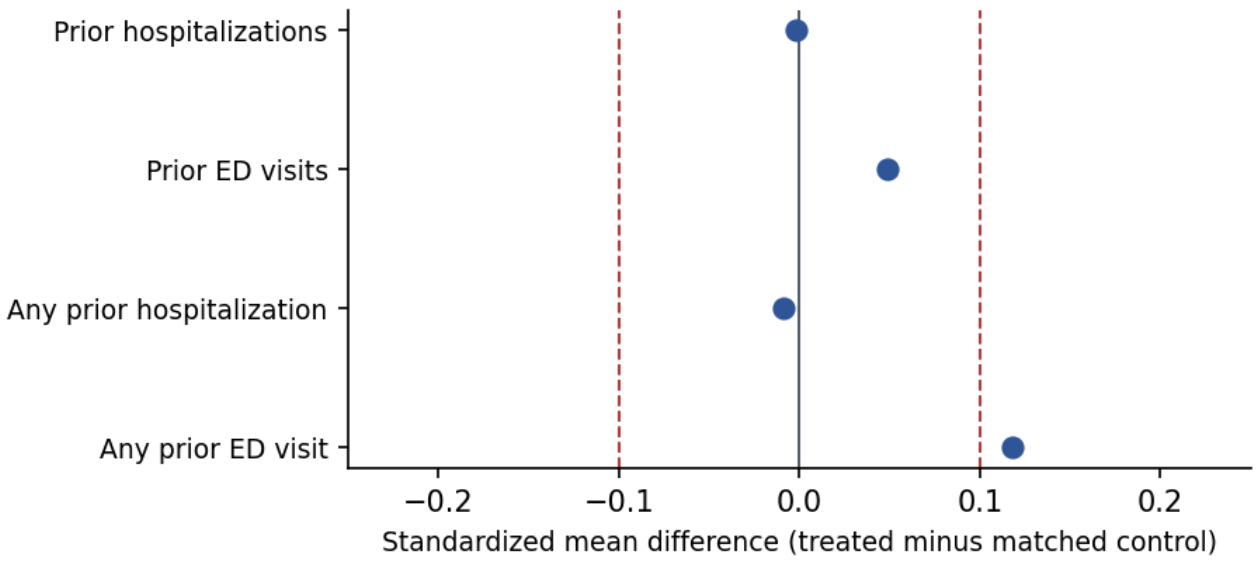

Figure A3: Balance on deep pre-2015 history. Standardized mean differences between matched treated patients and their weighted controls on hospitalizations and emergency visits recorded over 2009-2015, before any index month and outside the six-month window the design matches on. Dashed lines mark plus or minus 0.10; about 94 percent of patients link to this earlier administrative history.

Appendix tables

Balance and support

Table A1: Full descriptive statistics and baseline balance in the matched risk-set sample

Characteristic	Matched treated	Matched controls	Unmatched treated	SMD
Patients	136,113	136,113 (weighted)	97,719	–
Unique control patients	–	50,263	–	–
Female	76.6%	76.6%	63.6%	0.000
Age	58.6	58.5	56.9	0.003
Age 75+	15.5%	15.5%	17.5%	0.000
Index diagnosis R52 pain	37.4%	37.4%	22.7%	0.000
Index diagnosis M54 dorsalgia	19.7%	19.7%	12.6%	0.000
Index diagnosis C50 breast cancer	2.8%	2.8%	2.1%	0.000
Department 05 Antioquia	43.7%	43.7%	23.0%	0.000
Department 11 Bogota	28.7%	28.7%	20.2%	0.000
Department 76 Valle del Cauca	5.5%	5.5%	9.8%	0.000
Pre-6m recorded value	579,997	556,884	–	0.011
Recent 3m pre-index recorded value	308,501	290,494	–	0.013
Recorded value in month -1	112,439	102,497	–	0.013
Pre-6m consult count	6.51	6.30	–	0.032
Pre-6m procedure count	7.93	7.50	–	0.038
Active pre-history months	3.710	3.560	–	0.089
Diagnosis groups in pre-period	2.472	2.396	–	0.052
Pre-6m cancer diagnosis flag	13.6%	11.3%	–	0.069
Pre-6m pain diagnosis flag	52.2%	52.3%	–	-0.003
Pre-6m renal diagnosis flag	5.5%	3.2%	–	0.113

Table A2: Matching support, control reuse, and dual roles

Stage	Measure	Count	Share
Exposure scan	CUPS 890243 events, 2015-2021	684,529	–
Exposure scan	Unique patients with CUPS 890243	307,990	–
Exposure scan	CUPS 890343 follow-up/control events	398,900	–
Exposure scan	Old 2008-2015 exact-code matches for 890243/890343	0	–
Main sample	Eligible treated patients	233,832	100.0% of eligible
Main sample	Matched treated patients	136,113	58.2% of eligible
Main sample	Unmatched eligible treated patients	97,719	41.8% of eligible
Main sample	Matched control links	319,747	–
Main sample	Unique matched control patients	50,263	–
Main sample	Control links per unique control patient	6.4	–
Main sample	Control patients later matched as treated	16,548	32.9% of controls

Stage	Measure	Count	Share
Support restriction	No same calendar-month/dx/sex/department/age-cell control	68,738	70.3% of unmatched

Full pre-period lead diagnostics

Table A3: Monthly recorded-value lead estimates.

Relative month	Recorded-value lead (COP)	SE
-6	-2,102	2,072
-5	+351	2,024
-4	-4,699	2,051
-3	-409	2,100
-2	+769	2,180
-1	+6,090	2,279

Notes: Patient-level standard errors. The leads are within 6,090 COP of zero; the month -1 lead runs against the estimated decline if it reflects a persistent level difference. All six are within the clean-run criterion of about 9,000 COP.

Placebo construction and results

The placebo runs re-execute the complete pipeline after shifting every patient's first `890243` date earlier by 10 months (Placebo A) or 14 months (Placebo B), in both the treated-cohort file and the control-eligibility file. In the shifted world, no one receives any specialist consultation during the outcome window, control eligibility uses the shifted dates (so pseudo-controls are even-later-treated), and pseudo-treated patients are restricted to those with a recorded consultation in the pseudo-index month so that both arms remain anchored on an encounter, as in the real design. All matching bins, screens, caps, winsorization, and estimation are identical to the published pipeline.

Table A4: Placebo summary.

Quantity	Placebo A (shift 10)	Placebo B (shift 14)	Real analysis
Pseudo/actual treated matched	51,697	41,582	136,113
Months 1-9 estimate (COP/pm)	-42,380	+1,329	-44,875
SE	3,799	3,901	3,974*
Months 0-9 estimate (COP/pm)	-40,690	-2,998	-41,200
Largest absolute monthly lead	66,819	47,413	6,090
Leads significant at 5%	6/6	3/6	2/6
Averaged pre-period "placebo"	0 (by construction)	0 (by construction)	0 (by construction)
True effect in window	0	0	unknown

Notes: *The real-analysis SE is the person-cluster bootstrap; placebo SEs are patient-level. Placebo A's spurious estimate coincides with catastrophic lead failure (its pseudo-baseline is contaminated by the pre-referral escalation episode); Placebo B's recorded-value estimate is statistically zero despite three significant leads, illustrating that lead failure is necessary but not sufficient for an artifact. The pseudo cohorts are smaller than and compositionally different from the real cohort (the shifts move many index months outside the encounter-anchored eligibility window), so the placebo runs test the pipeline, not the population. The averaged pre-period check cannot distinguish the corrupted run from the valid one; the monthly leads can.

Service-code drivers

Appendix Table A5 splits the headline estimate into its consultation and procedure components, with the total carried by the person-cluster bootstrap interval. Appendix Tables A6 and A7 then list the consultation and procedure codes that contribute most to those components, with treated and control post-period means.

Table A5: Component split of the months 1-9 matched timing estimate.

Outcome	Estimate	SE	95% CI	Unit
Consultation + procedure recorded value	-44,875	3,974*	[-52,995, -37,385]	COP per patient-month
Consultation recorded value	-12,698	1,038	[-14,731, -10,664]	COP per patient-month
Procedure recorded value	-32,177	1,221	[-34,570, -29,784]	COP per patient-month
Other (non-specialist) consultation value	-17,401	1,049	[-19,457, -15,345]	COP per patient-month
Pain-and-palliative consultation value	+4,417	83	[4,255, 4,580]	COP per patient-month

Notes: Estimates are matched differences-in-differences net of each matched set's average pre-period difference (months -6 to -1). *The total carries the person-cluster bootstrap standard error and interval; component rows carry patient-level unclustered standard errors. The consultation subcomponents are estimated from separate code-level scans and need not sum exactly to the consultation row.

Table A6: Consultation-code recorded-value drivers, months 1-9 after the index consult

Direction	CUPS	Service	Matched difference	Treated post	Control post
Increase	890602	Inpatient management: specialized medicine	3,488	8,774	2,684
Increase	890243	First pain-and-palliative specialist consult	3,184	3,184	0
Increase	890343	Pain-and-palliative specialist follow-up	1,234	1,479	0
Increase	890503	Medical board: other health professional	274	530	104
Increase	890246	First consult: gastroenterology	177	123	296
Increase	890105	Home visit: nursing	157	1,326	643

Direction	CUPS	Service	Matched difference	Treated post	Control post
Increase	890232	First consult: breast/soft-tissue surgery	93	204	113
Increase	890101	Home visit: general medicine	75	878	771
Decrease	890202	First consult: other medical specialties	-3,694	1,221	6,369
Decrease	890302	Follow-up: other medical specialties	-3,670	745	11,089
Decrease	890201	First consult: general medicine	-2,980	1,442	4,694
Decrease	890301	Follow-up: general medicine	-2,718	1,393	5,921
Decrease	890206	First consult: nutrition/dietetics	-774	85	609
Decrease	890308	Follow-up: psychology	-508	72	560
Decrease	890207	First consult: optometry	-496	111	829

Table A7: Procedure-code recorded-value drivers, months 1-9 after the index consult

Direction	CUPS	Service	Matched difference	Treated post	Control post
Increase	933300	Hydraulic/hydrotherapy modalities	951	2,645	1,419
Increase	048301	Vertebral facet nerve anesthetic injection	814	1,012	112
Increase	053114	Regional sympathetic block	638	1,330	520
Increase	038200	Spinal root neurolysis	554	1,543	1,206
Increase	039101	Spinal canal anesthetic injection	354	386	24
Increase	053113	Continuous regional block	350	700	227
Increase	039001	Spinal injection family (CUPS 039001)	265	413	108
Increase	992300	Therapeutic application family (CUPS 992300)	228	305	139
Decrease	922444	Linear accelerator radiotherapy: IMRT	-3,108	3,568	5,433
Decrease	883230	MRI lumbosacral spine	-2,288	1,554	2,753
Decrease	992505	Multi-drug antineoplastic therapy	-1,232	2,409	3,507
Decrease	931001	Physical therapy: integral	-989	1,105	1,512
Decrease	922443	Linear accelerator radiotherapy: 3D-CRT	-959	2,551	2,610
Decrease	879601	PET-CT	-926	595	1,304
Decrease	931000	Physical therapy: integral (SOD variant)	-744	1,234	1,420
Decrease	883210	MRI cervical spine	-629	594	894

Exposure definition and phenotype

Table A8: CUPS codes used to define specialist pain-and-palliative contact.

CUPS	Use in analysis	Events	Patients	Matched patients	Cancer dx	Pain/MSK dx
890243	Main first-consult exposure	684,529	307,990	136,113	12.7%	75.5%
890343	Follow-up or control-contact marker	398,900	138,961	36,607	24.0%	64.8%
890443	Excluded interconsult code	165,600	78,803	3,048	42.9%	21.9%

Notes: Diagnosis shares are computed over each code's full patient set. The much higher cancer share for `890443` is why it stays outside the main exposure; an `890443`-based sensitivity analysis is left to future work.

Table A9: Clinical phenotype of matched treated patients

Measure	Matched treated share	Count
Index diagnosis in common pain/MSK groups	80.6%	109,737
Index diagnosis in cancer-related groups	6.7%	9,137
Any pre-period cancer diagnosis flag	13.6%	18,485
Any pre-period pain diagnosis flag	52.2%	71,049
Any pre-period musculoskeletal diagnosis flag	70.0%	95,287
Any pre-period serious non-cancer flag	53.1%	72,299
Age 75 or older	15.5%	21,090
Continuing specialist contact within 180 days	44.8%	60,915
No subsequent coded specialist follow-up observed	34.2%	46,520

Outcome definitions

Table A10: Outcome definitions and measurement limits

Outcome	RIPS source	Unit	Main limitation
Consultation + procedure service value	Consultas VALOR_NETO plus Procedimientos VALOR	Monthly COP sum, winsorized at positive 99.9 pct	Not patient payment or final EPS reimbursement; excludes drugs and full inpatient facility accounting; the two modules use different value fields
Consultation service value	Consultas VALOR_NETO	Monthly COP sum	Cannot separate negotiated payment from billed/support value; capitated contracts can record depressed values
Procedure service value	Procedimientos VALOR	Monthly COP sum	Procedure module may miss medicines, supplies, and non-procedure care inputs

Machine-learning review-list details

This appendix accompanies the review-list section. Appendix Table A11 reports the clean-protocol algorithm comparison, Appendix Table A12 the model against one-variable benchmarks, Appendix Table A13 the review-list rule menu (the green band requires an exact matched cell plus at least three of: three controls, match distance at or below the 75th percentile, at least six control links in the cell, and a strict history-match rule; yellow is exact support that misses that bar; red is no exact support), and Appendix Table A14 the composition audit of the held-out top decile. All numbers come from the re-validated protocol: model selection on the tuning cohorts, training-target winsorization computed on training rows only, and a single headline scoring of the test cohorts.

Table A11: Algorithm comparison under the clean protocol.

Model	Tune-stage top decile	Test top decile	Test share positive	Test lift
Gradient boosting, absolute error, engineered features (selected on tune)	316,177	602,210	80.0%	9.19
Gradient boosting, absolute error, original features	313,077	601,996	79.7%	9.19
Gradient boosting, squared error, default depth	311,944	587,511	76.4%	8.97
Extremely randomized trees, regressor, engineered	309,523	575,313	75.3%	8.78
Gradient boosting classifier, top-20% target, engineered	308,841	593,536	79.7%	9.06
Extremely randomized trees, classifier, top-20%, engineered	307,586	574,669	77.4%	8.77
Extremely randomized trees, regressor, original	307,342	578,443	73.5%	8.83
Gradient boosting classifier, top-20% target, original	307,238	600,686	79.7%	9.17
Gradient boosting, squared error, shallow regularized	306,819	622,643	79.1%	9.50
Extremely randomized trees, classifier, top-20%, original	304,934	568,959	76.9%	8.69
Gradient boosting classifier, top-10% target, original	274,448	577,009	76.7%	8.81
Gradient boosting classifier, top-10% target, engineered	272,393	585,966	78.5%	8.94

Notes: Selection uses the tune-stage column (most recent training cohorts); the test columns are reported for all twelve candidates as sensitivity only. Observed matched recorded-value reductions are in COP per patient-month over months 1-9; lift divides the top-decile mean by the test-cohort mean of 65,510 COP. This table uses the sweep harness's decile assignment, which differs from the publication exhibits' assignment by one patient at decile boundaries (602,210 here versus 600,982 elsewhere for the selected model). The narrow test range (569,000-623,000) shows the ranking value is not specific to one algorithm. Note that the model with the best test value ranked ninth on the tuning cohorts: choosing it would have meant selecting on the test set.

Table A12: The model against transparent benchmarks in the held-out cohorts.

Ranking rule	Top-decile mean observed matched reduction	Lift vs test mean	Share positive
Gradient-boosted model (392 features)	600,982	9.2	80.0%
Rank by 12-month pre-index spending	588,619	9.0	77.7%
Rank by recent 3-month spending	399,051	6.1	72.9%
Rank by month -1 spending	208,343	3.2	66.3%

Notes: All rules rank the same 20,388 test patients; reductions are observed matched recorded-value reductions in COP per patient-month over months 1-9, and positive values mean lower recorded spending than matched controls. Lift is the ratio of the rule's top-decile mean to the test-cohort mean.

Table A13: Review-list rules in the held-out test cohorts.

Review-list rule	Flags per month	Share with positive reduction	Mean observed reduction	Median
Top predicted decile, all matched support	127	80.0%	600,982	347,745
Top predicted decile, green support only	36	82.5%	436,015	299,482
Green support, predicted reduction > 0	346	64.4%	63,674	43,662
All matched support, predicted reduction > 0	805	62.8%	116,630	48,256

Notes: Flags per month average over the 16 test months; monthly cohort sizes vary within the window, so per-month counts range widely. The two "predicted reduction > 0" rows are reference points only: the calibration analysis shows the score's zero crossing is uninformative, so operational rules should be quantile-based. The strictest evidence rule (green-only top decile) has the highest hit rate but a lower mean than the unrestricted top decile because the largest measured reductions sit in the thin-support (yellow) band, the same pattern discussed in the main text.

Table A14: Composition audit of the held-out top decile (selected groups).

Dimension	Group	Share of cohort	Share of top decile	Representation ratio
Sex	Female	74.6%	71.4%	0.96
Sex	Male	25.4%	28.6%	1.13
Department	Bogota (11)	38.9%	32.9%	0.84
Department	Antioquia (05)	17.7%	17.9%	1.01
Department	Valle del Cauca (76)	10.3%	14.8%	1.44
Department	Santander (68)	7.9%	10.1%	1.28
Department	Risaralda (66)	4.5%	3.7%	0.83
Department	Bolivar (13)	2.3%	1.0%	0.43
Health plan	EPS037	25.3%	28.0%	1.11

Dimension	Group	Share of cohort	Share of top decile	Representation ratio
Health plan	EPS017	9.1%	11.0%	1.20
Health plan	EPS008	10.0%	6.8%	0.68
Health plan	EPS010	8.7%	3.0%	0.35

Notes: The representation ratio divides each group's share of the top decile by its share of the test cohort. Because the score is driven by recorded baseline spending, groups with suppressed recorded utilization, whether from access barriers or plan-specific reporting, are mechanically under-flagged; flags should not affect practice unless this audit is refreshed and reviewed.

Falsification outcomes, hidden-bias sensitivity, and the conditioning ladder

This appendix carries the detail behind the falsification and sensitivity results in the design tests. Table A15 reports the full monthly series for the two negative-control outcomes, estimated with exactly the published machinery (the same matched links, inverse-distance weights, and pre-period demeaning), with monthly cells capped at each line's own 99.9th percentile of positive values.

Table A15: Negative-control outcomes: monthly matched timing estimates.

Relative month	Dental DID	SE	Optometry DID	SE
-6	+9.6	13.7	-233.4	118.3
-5	+32.5	19.2	-35.4	142.9
-4	-26.1	23.6	+57.4	133.7
-3	-13.1	13.8	+133.6	51.6
-2	-39.0	20.1	+141.3	122.1
-1	+36.2	13.2	-63.4	78.1
0 (index)	-274.6	25.1	-223.5	132.6
1	+34.2	33.8	-599.2	234.6
2	+8.5	26.0	-28.2	132.0
3	+19.7	14.3	+120.8	58.0
4	+85.0	13.3	-876.9	177.5
5	+1.4	15.0	-549.8	154.0
6	-12.8	21.1	-505.2	145.6
7	-17.7	20.5	-2,313.1	267.4
8	+63.7	18.6	-210.4	107.8
9	+5.2	20.1	-131.2	96.9
Months 1-9	**+20.8**	10.8	**-565.9**	72.6

Notes: COP per patient-month; patient-level unclustered standard errors. Dental = CUPS 890203/890303 (first and follow-up general-dentistry consultations); optometry = CUPS 890207/890307. The index-month dips are mechanical (the specialist visit crowds out same-month elective contact). The headline outcome's months 1-9 estimate on the same links is -44,875.

The hidden-bias analysis treats each matched treated patient's months 1-9 demeaned contrast as one matched-set difference and applies Rosenbaum-style bounds: under hidden bias Gamma, the probability that a set's difference is negative is bounded by Gamma/(1+Gamma) when the true

effect is zero, and one-sided upper-bound p-values follow from normal approximations to the sign and signed-rank statistics (ties at zero dropped). In the full sample, 57.9 percent of the 135,946 nonzero contrasts are negative; the sign test rejects "no effect" at 5 percent up to Gamma = 1.37 and the signed-rank test up to Gamma = 1.48. In the pain/MSK core the bounds are 1.39 and 1.53. The rank-based tests are conservative relative to the mean-based estimate: monthly billing is noisy at the patient level, and the tests give no extra credit to the large declines that drive the average.

The conditioning ladder quantifies how much of the level selection the history matching absorbs. Against all candidates in their eligibility cells (same calendar month, diagnosis group, sex, department, and age bin, with no history screens and equal weights), matched treated patients run 51,587 COP per month cheaper in the pre-period and the cell-pool difference-in-differences is -11,798; the cell pool mixes in non-escalating patients whose flat trajectories do not represent the design's counterfactual, which is why the design screens them out. Against the chosen matches the pre-period gap is +3,852 (treated slightly more expensive), so the history screens remove 92.5 percent of the level selection and the residual biases the estimate toward zero; the post-period difference without any demeaning is -41,023, against the published -44,875. A fourth probe sought supply-driven timing variation: department-month first-consult throughput, residualized on department and calendar-month effects, predicts essentially none of the realized control delays (within R-squared 0.0005), so claims-derived congestion offers no usable first stage and quasi-experimental leverage requires external capacity data.

References

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, *74*, 235–267. https://doi.org/10.1111/j.1468-0262.2006.00655.x

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685–725. https://doi.org/10.1146/annurev-economics-080217-053433

Bakitas, M. A., Tosteson, T. D., Li, Z., Lyons, K. D., Hull, J. G., Li, Z., Dionne-Odom, J. N., Frost, J., Dragnev, K. H., Hegel, M. T., Azuero, A., & Ahles, T. A. (2015). Early versus delayed initiation of concurrent palliative oncology care: Patient outcomes in the ENABLE III randomized controlled trial. *Journal of Clinical Oncology*, *33*, 1438–1445. https://doi.org/10.1200/JCO.2014.58.6362

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, *119*, 249–275. https://doi.org/10.1162/003355304772839588

Borusyak, K., Jaravel, X., & Spiess, J. (2024). Revisiting event-study designs: Robust and efficient estimation. *Review of Economic Studies*, *91*, 3253–3285. https://doi.org/10.1093/restud/rdae007

Brumley, R., Enguidanos, S., Jamison, P., Seitz, R., Morgenstern, N., Saito, S., McIlwane, J., Hillary, K., & Gonzalez, J. (2007). Increased satisfaction with care and lower costs: Results of a randomized trial of in-home palliative care. *Journal of the American Geriatrics Society*, *55*, 993–1000. https://doi.org/10.1111/j.1532-5415.2007.01234.x

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*, 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*, 317–372. https://doi.org/10.3368/jhr.50.2.317

Congreso de la República de Colombia. (1993). Ley 100 de 1993, por la cual se crea el sistema de seguridad social integral. Diario Oficial No. 41.148.

Congreso de la República de Colombia. (2012). Ley 1581 de 2012, por la cual se dictan disposiciones generales para la protección de datos personales. Diario Oficial No. 48.587.

Congreso de la República de Colombia. (2014). Ley 1733 de 2014 (Ley Consuelo Devis Saavedra), mediante la cual se regulan los servicios de cuidados paliativos. Diario Oficial No. 49.268.

de Meijer, C., Koopmanschap, M., Bago d'Uva, T., & van Doorslaer, E. (2011). Determinants of long-term care spending: Age, time to death or disability? *Journal of Health Economics*, *30*, 425–438. https://doi.org/10.1016/j.jhealeco.2010.12.010

Felder, S., Werblow, A., & Zweifel, P. (2010). Do red herrings swim in circles? Controlling for the endogeneity of time to death. *Journal of Health Economics*, *29*, 205–212. https://doi.org/10.1016/j.jhealeco.2009.11.014

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*, 1189–1232. https://doi.org/10.1214/aos/1013203451

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, *225*, 254–277. https://doi.org/10.1016/j.jeconom.2021.03.014

Guerrero, R., Gallego, A. I., Becerril-Montekio, V., & Vásquez, J. (2011). Sistema de salud de Colombia. *Salud Pública de México*, *53*, s144–s155.

Hade, E. M., Nattino, G., Frey, H. A., & Lu, B. (2020). Propensity score matching for treatment delay effects. *Statistical Methods in Medical Research*, *29*, 695–708. https://doi.org/10.1177/0962280219877908

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, *64*, 605–654. https://doi.org/10.2307/2971733

Kavalieratos, D., Corbelli, J., Zhang, D., Dionne-Odom, J. N., Ernecoff, N. C., Hanmer, J., Hoydich, Z. P., Ikejiani, D. Z., Klein-Fedyshin, M., Zimmermann, C., Morton, S. C., Arnold, R. M., Heller, L., & Schenker, Y. (2016). Association between palliative care and patient and caregiver outcomes: A systematic review and meta-analysis. *JAMA*, *316*, 2104–2114. https://doi.org/10.1001/jama.2016.16840

Kelley, A. S., & Morrison, R. S. (2015). Palliative care for the seriously ill. *New England Journal of Medicine*, *373*, 747–755. https://doi.org/10.1056/NEJMra1404684

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, *105*, 491–495. https://doi.org/10.1257/aer.p20151023

Knaul, F. M., Farmer, P. E., Krakauer, E. L., De Lima, L., Bhadelia, A., Jiang Kwete, X., Arreola-Ornelas, H., Gómez-Dantés, O., Rodriguez, N. M., Alleyne, G. A. O., Connor, S. R., Hunter, D. J., Lohman, D., Radbruch, L., Sáenz Madrigal, M. del R., Atun, R., Foley, K. M.,

Frenk, J., Jamison, D. T., & Rajagopal, M. R. (2018). Alleviating the access abyss in palliative care and pain relief—An imperative of universal health coverage: The Lancet Commission report. *Lancet*, *391*, 1391–1454. https://doi.org/10.1016/S0140-6736(17)32513-8

Li, Y. P., Propert, K. J., & Rosenbaum, P. R. (2001). Balanced risk set matching. *Journal of the American Statistical Association*, *96*, 870–882. https://doi.org/10.1198/016214501753208573

Lipsitch, M., Tchetgen Tchetgen, E., & Cohen, T. (2010). Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, *21*, 383–388. https://doi.org/10.1097/EDE.0b013e3181d61eeb

Lu, B. (2005). Propensity score matching with time-dependent covariates. *Biometrics*, *61*, 721–728. https://doi.org/10.1111/j.1541-0420.2005.00356.x

May, P., Garrido, M. M., Cassel, J. B., Kelley, A. S., Meier, D. E., Normand, C., Smith, T. J., Stefanis, L., & Morrison, R. S. (2015). Prospective cohort study of hospital palliative care teams for inpatients with advanced cancer: Earlier consultation is associated with larger cost-saving effect. *Journal of Clinical Oncology*, *33*, 2745–2752. https://doi.org/10.1200/JCO.2014.60.2334

May, P., Garrido, M. M., Cassel, J. B., Kelley, A. S., Meier, D. E., Normand, C., Smith, T. J., & Morrison, R. S. (2017). Cost analysis of a prospective multi-site cohort study of palliative care consultation teams for adults with advanced cancer: Where do cost savings from palliative care come from? *Palliative Medicine*, *31*, 378–386. https://doi.org/10.1177/0269216317690098

May, P., Normand, C., Cassel, J. B., Del Fabbro, E., Fine, R. L., Menz, R., Morrison, C. A., Penrod, J. D., Robinson, C., & Morrison, R. S. (2018). Economics of palliative care for hospitalized adults with serious illness: A meta-analysis. *JAMA Internal Medicine*, *178*, 820–829. https://doi.org/10.1001/jamainternmed.2018.0750

Ministerio de Salud y Protección Social. (2022). Registro Individual de Prestación de Servicios de Salud (RIPS) [Data set]. Sistema Integrado de Información de la Protección Social (SISPRO). https://www.sispro.gov.co/ [VERIFY: confirm dataset year/version and URL]

Morrison, R. S., Penrod, J. D., Cassel, J. B., Caust-Ellenbogen, M., Litke, A., Spragens, L., & Meier, D. E. (2008). Cost savings associated with US hospital palliative care consultation programs. *Archives of Internal Medicine*, *168*, 1783–1790. https://doi.org/10.1001/archinte.168.16.1783

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*, 87–106. https://doi.org/10.1257/jep.31.2.87

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future: Big data, machine learning, and clinical medicine. *New England Journal of Medicine*, *375*, 1216–1219. https://doi.org/10.1056/NEJMp1606181

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*, 447–453. https://doi.org/10.1126/science.aax2342

Pastrana, T., & De Lima, L. (2022). Palliative care in Latin America. *Journal of Pain and Symptom Management*, *63*, 33–41. https://doi.org/10.1016/j.jpainsymman.2021.07.020

Pastrana, T., De Lima, L., Sánchez-Cárdenas, M., Van Steijn, D., Garralda, E., Pons-Izquierdo, J. J., & Centeno, C. (2020). *Atlas de cuidados paliativos en Latinoamérica 2020* (2nd ed.). IAHPC Press.

Quill, T. E., & Abernethy, A. P. (2013). Generalist plus specialist palliative care: Creating a more sustainable model. *New England Journal of Medicine*, *368*, 1173–1175. https://doi.org/10.1056/NEJMp1215620

Rambachan, A., & Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, *90*, 2555–2591. https://doi.org/10.1093/restud/rdad018

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4757-3692-2

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, *4*, 305–322. https://doi.org/10.1257/aeri.20210236

Sanchez-Cardenas, M. A., León, M. X., Rodríguez-Campos, L. F., Vargas-Escobar, L. M., Cabezas, L., Tamayo-Díaz, J. P., Cañón Piñeros, A., Mantilla-Manosalva, N., & Fuentes-Bermúdez, G. P. (2024). Accessibility to palliative care services in Colombia: An analysis of geographic disparities. *BMC Public Health*, *24*, 1659. https://doi.org/10.1186/s12889-024-19132-2

Sanders, J. J., Temin, S., Ghoshal, A., Alesi, E. R., Ali, Z. V., Chauhan, C., Cleary, J. F., Epstein, A. S., Firn, J. I., Jones, J. A., Litzow, M. R., Lundquist, D., Mardones, M. A., Nipp, R. D., Rabow, M. W., Rosa, W. E., Zimmermann, C., & Ferrell, B. R. (2024). Palliative care for patients with cancer: ASCO guideline update. *Journal of Clinical Oncology*, *42*, 2336–2357. https://doi.org/10.1200/JCO.24.00542

Seow, H., Barbera, L. C., McGrail, K., Burge, F., Guthrie, D. M., Lawson, B., Chan, K. K. W., Peacock, S. J., & Sutradhar, R. (2022). Effect of early palliative care on end-of-life health care costs: A population-based, propensity score–matched cohort study. *JCO Oncology Practice*, *18*, e183–e192. https://doi.org/10.1200/OP.21.00299

Seshamani, M., & Gray, A. M. (2004). Ageing and health-care expenditure: The red herring argument revisited. *Journal of Health Economics*, *23*, 217–235. https://doi.org/10.1016/j.jhealeco.2003.08.004

Sheridan, P. E., LeBrett, W. G., Triplett, D. P., Roeland, E. J., Bruggeman, A. R., Yeung, H. N., & Murphy, J. D. (2021). Cost savings associated with palliative care among older adults with advanced cancer. *American Journal of Hospice and Palliative Medicine*, *38*, 1250–1257. https://doi.org/10.1177/1049909120986800

Sleeman, K. E., de Brito, M., Etkind, S., Nkhoma, K., Guo, P., Higginson, I. J., Gomes, B., & Harding, R. (2019). The escalating global burden of serious health-related suffering: Projections to 2060 by world regions, age groups, and health conditions. *Lancet Global Health*, *7*, e883–e892. https://doi.org/10.1016/S2214-109X(19)30172-X

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*, 175–199. https://doi.org/10.1016/j.jeconom.202

Temel, J. S., Greer, J. A., Muzikansky, A., Gallagher, E. R., Admane, S., Jackson, V. A., Dahlin, C. M., Blinderman, C. D., Jacobsen, J., Pirl, W. F., Billings, J. A., & Lynch, T. J. (2010). Early palliative care for patients with metastatic non-small-cell lung cancer. *New England Journal of Medicine*, *363*, 733–742. https://doi.org/10.1056/NEJMoa1000678

Vargas-Escobar, L. M., Sánchez-Cárdenas, M. A., Guerrero-Benítez, A. C., Suarez-Prieto, V. K., Moreno-García, J. R., Cañón Piñeros, Á. M., Rodríguez-Campos, L. F., & León-Delgado, M. X. (2022). Barriers to access to palliative care in Colombia: A social

mapping approach involving stakeholder participation. *Inquiry*, *59*, 00469580221133217. https://doi.org/10.1177/00469580221133217

Werblow, A., Felder, S., & Zweifel, P. (2007). Population ageing and health care expenditure: A school of 'red herrings'? *Health Economics*, *16*, 1109–1126. https://doi.org/10.1002/hec.1213

Wong, A., van Baal, P. H. M., Boshuizen, H. C., & Polder, J. J. (2011). Exploring the influence of proximity to death on disease-specific hospital expenditures: A carpaccio of red herrings? *Health Economics*, *20*, 379–400. https://doi.org/10.1002/hec.1597

Zimmermann, C., Swami, N., Krzyzanowska, M., Hannon, B., Leighl, N., Oza, A., Moore, M., Rydall, A., Rodin, G., Tannock, I., Donner, A., & Lo, C. (2014). Early palliative care for patients with advanced cancer: A cluster-randomised controlled trial. *Lancet*, *383*, 1721–1730. https://doi.org/10.1016/S0140-6736(13)62416-2

Zweifel, P., Felder, S., & Meiers, M. (1999). Ageing of population and health care expenditure: A red herring? *Health Economics*, *8*, 485–496. https://doi.org/10.1002/(SICI)1099-1050(199909)8:6%3C485::AID-HEC461%3E3.0.CO;2-4