

# HEDG

HEALTH, ECONOMETRICS AND DATA GROUP

---

THE UNIVERSITY *of York*

WP 26/16

## Beyond Average: Who Benefits from Integrated Care? Heterogeneous Treatment Effects on Healthcare Utilization and Costs

Michael Möller; Eva-Maria Wild and Vanessa Röss

August 2026

<http://www.york.ac.uk/economics/postgrad/herc/hedg/wps/>

# **Beyond Average: Who Benefits from Integrated Care? Heterogeneous Treatment Effects on Healthcare Utilization and Costs**

Authors: Michael Möller<sup>1</sup>, Eva-Maria Wild<sup>2</sup>, Vanessa Ress<sup>2</sup>

## **Abstract**

Although integrated care has attracted increasing attention, there is limited causal evidence on heterogeneous treatment effects across patient groups. This gap is important for health-economic evaluation as average effects can miss clinically and economically relevant variation in healthcare utilization and cost responses. For resource-intensive integrated care programs, the economic value may depend on which patient groups experience changes in hospital-based care, outpatient utilization, and healthcare costs.

We evaluate an integrated care initiative in a socially deprived urban area in Germany using claims data for 199,448 insured individuals observed from 2015 to 2024. Because enrollment in the program is staggered and selective, we estimate heterogeneous treatment effects using a staggered Difference-in-Differences Causal Forest, comparing enrolled individuals to not-yet-enrolled and never-enrolled counterparts and flexibly recovering how treatment effects vary with pre-treatment patient characteristics.

Heterogeneity in treatment effects is most pronounced for inpatient costs, hospital outpatient visits, and ambulatory care-sensitive condition admissions. Estimated reductions in inpatient costs and ambulatory care-sensitive condition admissions are concentrated among patients with higher pre-treatment hospital utilization, a history of avoidable hospitalizations, and elevated inpatient costs. Hospital outpatient visits increase more among patients with greater baseline

---

<sup>1</sup> University of Hamburg, Hamburg Center for Health Economics, Esplanade 36, 20354 Hamburg, Germany  
E-mail: [michael.moeller-2@uni-hamburg.de](mailto:michael.moeller-2@uni-hamburg.de)

<sup>2</sup> University of Siegen, School of Economic Disciplines, Siegen, Germany

ambulatory engagement, and individuals with limited prior acute-care use exhibit smaller effects.

These findings indicate that the economic implications of integrated care are contingent on how utilization and cost effects are distributed across patient groups, particularly with respect to prior acute-care use, healthcare costs, and care intensity. Health-economic evaluations of integrated care should incorporate heterogeneous treatment effect estimation to distinguish between groups for whom such interventions may reduce hospital-based care and those for whom the primary effect is increased ambulatory engagement.

*Keywords:* Causal machine learning, Heterogeneous treatment effects, Claims data, Integrated care, Healthcare utilization, Healthcare costs

## 1 Introduction

Healthcare systems are confronted with increasing care demands, limited resources, and persistent disparities in access. Demographic aging, multimorbidity, and social disadvantage drive the need for coordinated care services, while fiscal constraints intensify these challenges (OECD, 2024). Integrated care (IC) models address these issues by offering accessible advice, navigation, and coordination across providers and sectors, aiming to enhance continuity of care, identify unmet needs, and reduce avoidable hospital utilization (OECD, 2023). However, IC models also require staff time, provider coordination, infrastructure, and financing. In this paper, we examine how IC influences healthcare utilization and costs across patient groups, as its economic value may depend on whether the effects are concentrated among patients with higher baseline utilization of costly hospital-based or avoidable acute care (Wammes et al., 2018; Kent et al., 2020; Rocks et al., 2020; Quinton et al., 2021).

Economic evaluations typically report average effects, which may mask important variation in program impacts across patients with differing baseline morbidity, prior utilization, and costs. Modest or imprecise average effects can reflect large reductions in hospital-based care for some patients alongside limited effects, or increased outpatient use and costs, for others. Estimating heterogeneous treatment effects (HTE) is therefore necessary for the economic evaluation of IC (Basu and Meltzer, 2007; Grutters et al., 2013; Basu et al., 2014; Espinoza et al., 2014; Basu et al., 2019). This study advances the health-economic evaluation of IC by estimating HTEs directly, using a staggered Difference-in-Differences Causal Forest (DiDCF) that combines cohort-event-time DiD identification with a causal forest estimating conditional average treatment effects on the treated (CATTs), and by assessing whether this heterogeneity is systematically related to observable baseline patient profiles. In doing so, we contribute new evidence to a literature that has generally reported modest or mixed average effects for IC and



related care-coordination interventions (Goldzahl et al., 2022; Lippi Bruni et al., 2023; Ress and Wild, 2024; Bancalari et al., 2026).

This combination of design features is required by the setting. IC enrollment is staggered and selective, and treatment effects are plausibly heterogeneous across a diverse target population. Under these conditions, conventional two-way fixed-effects (TWFE) estimators are known to be biased (Chaisemartin and D’Haultfœuille, 2020; Goodman-Bacon, 2021). Modern staggered-DiD estimators address this bias by comparing each cohort only to valid untreated observations (Callaway and Sant’Anna, 2021; Borusyak et al., 2024). These estimators, however, are designed to recover average or, at most, prespecified subgroup effects, and do not characterize how treatment effects vary across a high-dimensional set of pre-treatment patient characteristics.

Identifying which patients benefit therefore poses a separate problem from correcting for staggered adoption: fixing subgroups in advance risks data-mining, while the relevant patient profile is unlikely to be captured by a single prespecified characteristic. Answering this instead requires a method that flexibly recovers heterogeneity from a high-dimensional set of pre-treatment claims histories without relying on ad hoc subgroup definitions. This connects recent advances in causal-ML-based HTE estimation to the question of whether program value can be assessed from average effects alone, or requires evidence on who benefits and who does not. (Marafino et al., 2020; Kreif et al., 2022; Glynn et al., 2024; Shah and Kreif, 2025; Shah et al., 2026).

We apply this framework to an integrated care initiative in a socially deprived urban area in Germany, which seeks to reduce health inequities related to social circumstances, improve health opportunities for the population, enhance patient experience and satisfaction, and optimize resource use through community-based advice, navigation, and coordination with

medical and social care providers (Ress and Wild, 2024). Community-based care coordination targeted at socioeconomically disadvantaged populations is a widely used policy instrument internationally, and whether its economic value depends on patients' prior care histories is a question relevant beyond this specific program (Macinko et al., 2010; Bailey and Goodman-Bacon, 2015; Baxter et al., 2018; OECD, 2023; Bancalari et al., 2026). The intervention may reduce avoidable hospital admissions, emergency department use, ambulatory care-sensitive condition (ACSC) admissions, and inpatient costs by improving coordination and care transitions. Conversely, it may increase outpatient contacts, monitoring, social care contacts, or short-term costs if previously unmet needs are identified and addressed. This study empirically examines changes in utilization and costs across care sectors.

We use administrative insurance claims linked to IC program information to construct a person-quarter panel covering 2015 to 2024. The primary analysis focuses on individuals residing in the program's target area. Treatment is defined as formal enrollment in the IC program, which occurs at the patient level at different points during the observation period. Outcomes include healthcare utilization and costs across sectors, such as hospital admissions, emergency admissions, ACSC outcomes, and non-hospital ambulatory care. Pre-treatment information on prior utilization and costs, morbidity-related care needs, sickness absence, residential location, and prior participation in disease-management and other IC programs is used to adjust for differences before enrollment and to estimate treatment effects across patient profiles. We compare outcome changes among enrolled individuals with those among not-yet-enrolled and never-enrolled individuals observed over the same calendar periods to estimate cohort-event-time CATTs, with identification relying on conditional parallel trends and common support (Gavrilova et al., 2025).

The findings show that average effects provide only a partial summary of the program's impact. Inpatient utilization changes minimally on average, while non-hospital ambulatory utilization

and costs increase more noticeably. However, these average patterns obscure significant differences across patient groups. Heterogeneity is most pronounced for inpatient costs, hospital outpatient visits, and ACSC admissions. Some patients experience sizable reductions in inpatient costs and ACSC admissions, while hospital outpatient visits increase more for others, particularly those with greater baseline ambulatory engagement. Across outcomes, estimated reductions are concentrated among patients with greater pre-period hospital use, ACSC history, and inpatient costs, and patients with little prior acute-care use generally show smaller estimated responses. These results suggest that the value of IC cannot be assessed solely from average treatment effects and depends on how utilization and cost effects are distributed across patient groups.

The remainder of the paper is structured as follows. Section 2 summarizes prior evidence on the evaluation of integrated care and HTE estimation. Section 3 introduces the institutional setting. Section 4 describes the data and sample construction. Section 5 presents the empirical strategy, and Section 6 reports the results. Finally, Section 7 discusses the findings and concludes the paper.

## **2 Literature Review**

### *2.1 Integrated care and community-based navigation*

IC, community health teams, and coordinated primary-care models are designed to improve care coordination, access, and continuity across sectors (Khatri et al., 2023a; Khatri et al., 2023b). Evaluations of these interventions indicate that effects on healthcare utilization vary by institutional context, intervention design, target population, and outcome. In Germany, studies have examined mechanisms including social prescribing, community health advice, changes in primary-care demand, and impacts on hospital admissions, emergency care, ambulatory visits, and costs (Golubinski et al., 2020; Wild et al., 2022; Ress and Wild, 2024). International studies

similarly report mixed average effects on mortality, ambulatory-care-sensitive hospitalizations, emergency service use, outpatient care, hospital utilization, and costs. (Macinko et al., 2010; Bailey and Goodman-Bacon, 2015; Goldzahl et al., 2022; Lippi Bruni et al., 2023; Bancalari et al., 2026). Most existing evidence addresses average effects with less attention to systematic variation in effects across patient groups.

## *2.2 HTE and economic evaluation*

A growing health economics and health services literature uses causal machine learning to estimate HTEs and to describe variation in treatment response. This literature shows that average effects can conceal clinically and economically important differences across patient groups. Applications to health insurance, social policy, and healthcare delivery show that treatment responses often vary by socioeconomic characteristics, prior utilization, baseline risk, and supply-side conditions (Marafino et al., 2020; Kreif et al., 2022; Glynn et al., 2024; Denteh and Liebert, 2025; Shah and Kreif, 2025; Shah et al., 2026).

This literature is relevant to our study because evidence on average effects identifies whether outcomes such as hospital admissions, avoidable acute care, outpatient utilization, or healthcare costs change overall, but does not indicate whether these changes are concentrated within specific patient subgroups. In the context of IC, HTEs are plausible given variation in morbidity burden, prior utilization, baseline costs, care needs, and previous exposure to coordinated-care structures (Vuik et al., 2016; Quinton et al., 2021; Quinton et al., 2023).

## **3 Institutional Setting**

We examine the integrated care initiative *INVEST Billstedt/Horn*, implemented in the fourth quarter of 2017 in a socially deprived urban district of Hamburg, Germany. The intervention area comprises approximately 110,000 residents with low average income, high levels of

socioeconomic disadvantage, and increased health and care needs. Program services are universally available to residents and provided at no direct cost at the point of use.

The primary component of the program is a community health advice and navigation service delivered by community health nurses. These nurses offer accessible counseling, health information, referral support, and assistance with navigating medical and social care systems.

The intervention may affect healthcare use through several channels. Advice and navigation can reduce informational and administrative barriers to appropriate care; provider coordination may improve follow-up, medication management, and care transitions; and referral support may connect patients with ambulatory or social services before needs become urgent. At the same time, improved access may increase measured utilization and costs, especially in the short-term, if previously unmet needs are identified or monitoring becomes more intensive.

For evaluation purposes, the program generates individual-level enrollment at varying times during the observation period. Enrollment records can be linked to pre-enrollment claims data and to post-enrollment utilization and cost outcomes. Further details on the institutional setting are available in (Ress and Wild, 2024).

## **4 Data**

### *4.1 Data source and panel construction*

We analyze longitudinal administrative claims data from a large German statutory health insurer for the period 2015 to 2024. The dataset integrates insurance records, sector-specific claims, and program enrollment information. Insurance records include demographic characteristics, insurance status, employment status, and residential location, distinguishing between target districts and other areas of Hamburg. Claims records document healthcare utilization and insurer-reimbursed costs for inpatient care, hospital outpatient care, non-hospital

ambulatory care, prescription drugs, and long-term care. Program records identify individuals formally enrolled in the IC program and record the enrollment date.

We construct an unbalanced person-quarter panel, harmonizing all outcomes and time-varying covariates. Individuals contribute one observation for each quarter in which they are insured and observable in the administrative data. The full panel comprises 199,448 individuals and 7,029,739 person-quarter observations.

Treatment is defined as formal enrollment in the IC program, documented by a signed agreement and required for access to program services. Individuals can enroll through various entry channels, including referrals from participating providers, social care institutions, or direct contact with the program. Enrollment occurs at the individual level and at varying times during the observation period, resulting in staggered treatment adoption. Treated individuals are assigned to cohorts based on their first observed enrollment quarter, which is defined as the treatment start. Event time is measured relative to the enrollment quarter, with event time  $\ell = 0$  denoting the quarter of initial enrollment.

#### *4.2 Outcomes and covariates*

We analyze quarterly utilization and cost outcomes across healthcare sectors. Utilization outcomes comprise inpatient admissions, hospital outpatient visits, non-hospital ambulatory visits, emergency care, and ACSC admissions. Cost outcomes include sector-specific and total healthcare costs, measured from the insurer's perspective and aggregated to the person-quarter level.

To address potential misclassification, we construct two emergency-related outcomes. The first identifies emergency admissions based on the recorded reason for hospital admission. The second identifies admissions using primary inpatient ICD-10 diagnoses, following the approach of (Krämer et al., 2019)). ACSC admissions are identified using the core list in (Sundmacher et

al., 2015) and are based on inpatient diagnoses observed in the claims data. Further details on outcome construction are provided in Appendix A.

Covariates are constructed from pre-treatment claims and insurance histories. The main covariates include age, sex, residence area, four-quarter pre-enrollment histories of hospital admissions, outpatient visits, sector-specific costs, medication costs, long-term care costs, sickness absence, ACSC admissions, and prior participation in disease-management or other IC programs.

For treated individuals, pre-treatment covariates are measured over the four quarters preceding IC enrollment unless otherwise specified. For not-yet-treated and never-treated comparison observations, covariates are measured over the same four calendar quarters as for the treated cohort to which they are compared. When a cohort is compared with eligible untreated observations, covariates for both groups are constructed from the cohort-specific pre-enrollment window. This approach ensures that covariates used for adjustment and heterogeneity estimation are predetermined with respect to the treatment comparison.

#### *4.3 Sample construction*

For the main analysis, we restrict the full panel to individuals residing in Billstedt, Horn, and adjacent districts, corresponding to the catchment area of the IC initiative's access points (Figure A1). This restriction defines the population plausibly exposed to the program and its local provider network. We further restrict the main analysis panel to individuals aged 18 or older at the cohort-specific baseline quarter. The resulting main analysis panel comprise 50,187 individuals and 1,602,685 person-quarter observations.

During the study period, 4,336 individuals (8.6%) in our sample enrolled in the IC program. In the staggered difference-in-differences (DiD) design, enrolled individuals contribute untreated observations before their first enrollment quarter and treated observations thereafter.

Comparison observations are drawn dynamically from individuals not-yet enrolled in a given calendar quarter and from those never observed to enroll during the study period. This structure allows treated cohorts to be compared with not-yet-treated and never-treated individuals in the same calendar quarters, preserving the quarterly event-time structure around enrollment.

The final estimation samples can differ from the base population because not every cohort-event-time comparison is observable within the event-study window and because DiDCF estimation requires sufficient treated and untreated observations within common support. Estimation sample sizes and support diagnostics are reported in the results and Appendix A.

## **5 Empirical Framework**

We apply the DiDCF framework of (Gavrilova et al., 2025) to estimate HTEs under staggered treatment timing. This approach integrates a staggered DiD design with causal forests. The DiD component defines cohort-event-time comparisons between enrolled individuals and valid untreated comparison observations. The causal forest estimates variation in treatment effects as a function of pre-treatment patient characteristics within these comparisons. The primary estimand is the cohort-event-time conditional average treatment effect on the treated (CATT).

Appendix D provides full technical detail on the DiDCF method, including the orthogonal moment condition, the two-stage nuisance and forest estimation procedure, and the doubly robust aggregation of cell-level ATTs.

### *5.1 Staggered treatment timing and notation*

The empirical setting is a staggered treatment design in which individuals enroll in the IC program in different calendar quarters, while a subset remains untreated throughout the observation period. Enrollment occurs across 29 treatment cohorts, spanning from the fourth quarter of 2017 to the fourth quarter of 2024. Effects are estimated over an event-time window from four quarters prior to enrollment to eight quarters following enrollment.



Let  $c \in \{1, \dots, C\}$  denote treatment cohorts, where a cohort is the set of individuals who enroll in the program in the same calendar quarter, and  $C$  denotes the total number of cohorts. We use  $G_i$  to denote individual  $i$ 's enrollment quarter, with  $G_i = \infty$  for individuals who never enroll within the observed window. Cohort  $c$  enrolls in calendar quarter  $g_c$ . Its baseline period is denoted by  $b_c$ . In our main analysis,  $b_c = g_c - 1$ , so that the baseline is the quarter immediately before enrollment.

We use  $\ell$  for event time relative to enrollment,

$$\ell = t - g_c,$$

so that  $\ell = 0$  is the enrollment quarter and  $\ell = -1$  is the baseline quarter. The corresponding calendar quarter is  $t = g_c + \ell$ .

Because enrolled individuals may differ from comparison individuals in baseline morbidity, utilization, and costs, we do not directly compare post-enrollment outcome levels. Instead, we define outcome changes relative to the cohort-specific baseline period as

$$z_{i,c,\ell} = Y_{i,g_c+\ell} - Y_{i,b_c},$$

where  $z_{i,c,\ell}(1)$  and  $z_{i,c,\ell}(0)$  denote the baseline-differenced potential outcomes with and without IC enrollment. The baseline-differenced outcome captures the change in individual  $i$ 's outcome  $Y_i$  from the cohort-specific baseline  $b_c$  to event time  $\ell$ . It turns the comparison into a DiD comparison and measures how outcomes change for enrolled individuals relative to how outcomes change over the same calendar period for comparable individuals who remain untreated.

For cohort  $c$  at event time  $\ell$ , valid comparison observations are untreated individuals in the quarter  $g_c + \ell$ , that is, never-treated and not-yet-treated individuals. We denote this set by

$$\mathcal{U}_{c,\ell} = \{i: G_i = \infty \text{ or } G_i > g_c + \ell\}.$$

Already-treated individuals are not used as controls for later-treated cohorts. This restriction addresses the bias that arises when TWFE estimators implicitly compare newly treated individuals to already-treated individuals under staggered adoption, particularly when treatment effects are heterogeneous across cohorts, event times, or patient groups (Chaisemartin and D’Haultfœuille, 2020; Goodman-Bacon, 2021).

Within each cohort-event-time comparison  $(c, \ell)$ , we define

$$W_i^{c,\ell} = 1\{G_i = g_c\}$$

as an indicator for belonging to the enrolled cohort. Thus,  $W_i^{c,\ell} = 1$  for individuals enrolled in cohort  $c$  and  $W_i^{c,\ell} = 0$  for valid untreated comparison observations in  $\mathcal{U}_{c,\ell}$ .

The estimand of interest is the cohort-event-time CATT. For cohort  $c$ , event time  $\ell$ , and pre-treatment covariates  $x$ , we define

$$CATT(c, \ell, x) = E[z_{i,c,\ell}(1) - z_{i,c,\ell}(0) \mid W_i^{c,\ell} = 1, X_i = x].$$

Conditioning on  $W_i^{c,\ell} = 1$  defines the estimand for individuals in the enrolled cohort. The untreated potential outcome  $z_{i,c,\ell}(0)$  is counterfactual and is identified using eligible not-yet-treated and never-treated comparison observations subject to the conditional parallel trends assumption described below.

## 5.2 Identification assumptions

The plausibility of parallel trends in this setting rests on both statistical and institutional grounds. Not-yet-enrolled and never-enrolled individuals residing in the same catchment area face the same local provider network, the same regional secular shocks, and the same period-specific disruptions, including the COVID-19 pandemic, as enrolled individuals. Common time effects are therefore plausible by construction within the target-area sample. The more specific

threat to identification is not aggregate area-level shocks but individual-level events that trigger both enrollment and the outcome, discussed below.

The key identifying assumption underlying modern staggered and semiparametric DiD approaches is conditional parallel trends in untreated potential outcomes (Callaway and Sant’Anna, 2021). Conditional on pre-treatment covariates  $X_i$ , treated individuals in cohort  $c$  and valid comparison individuals untreated at event time  $\ell$  would have experienced the same average outcome change in the absence of enrollment:

$$E[z_{i,c,\ell}(0) \mid W_i^{c,\ell} = 1, X_i = x] = E[z_{i,c,\ell}(0) \mid W_i^{c,\ell} = 0, X_i = x].$$

This assumption does not require treated and comparison individuals to have the same baseline outcome levels. It requires that, after conditioning on observed pre-treatment histories, their untreated outcome changes are comparable.

We also require the assumptions of no anticipation, overlap, and no interference (Imbens and Rubin, 2015; Callaway and Sant’Anna, 2021). No anticipation excludes changes in outcomes prior to observed enrollment that are attributable to future program participation. Overlap requires the presence of comparable untreated observations for treated individuals within each cohort-event-time comparison. No interference requires that one individual’s potential outcomes are not directly affected by another individual’s enrollment status. We assess these assumptions using pre-treatment event-study estimates, placebo timing checks, overlap and balance diagnostics, and robustness analyses.

### *5.3 DiD Causal Forest estimation of heterogeneous effects*

The DiDCF estimates  $\tau_{c,\ell}(x)$ , the cohort-event-time CATT for individuals with pre-treatment covariates  $x$ , by comparing baseline-differenced outcomes between enrolled and untreated

individuals with similar pre-treatment covariates. The causal forest models how treatment-induced outcome changes vary as a function of baseline patient characteristics.

For each cohort-event-time comparison, the method uses two nuisance functions. The outcome regression,

$$m_{c,\ell}(X_i) = E[z_{i,c,\ell} \mid X_i],$$

predicts the baseline-differenced outcome from pre-treatment covariates. The propensity model,

$$p_{c,\ell}(X_i) = P(W_i^{c,\ell} = 1 \mid X_i),$$

predicts membership in the enrolled cohort within the cohort-event-time comparison. These two nuisance functions are used to residualize the outcome change and treatment indicator with respect to observed pre-treatment covariates. This orthogonalization step adjusts for differences between enrolled and untreated observations prior to estimation of treatment-effect heterogeneity by the causal forest (Chernozhukov et al., 2018; Athey et al., 2019).

The causal forest then constructs data-adaptive neighborhoods of individuals with similar pre-treatment characteristics. Observations with covariate profiles closer to a given  $x$  receive larger forest weights when estimating  $\tau_{c,\ell}(x)$ . While this approach is analogous to matching, the estimator is not a conventional matching estimator. Propensity scores are used for adjustment within the DiDCF score rather than to define a separate matched comparison sample. Comparability is addressed through the conditional DiD design, overlap restrictions, forest-based local weighting, and balance diagnostics (ROSENBAUM and Rubin, 1983; Athey et al., 2019).

The resulting orthogonalized estimator is designed to reduce sensitivity to errors in the nuisance models and, when nuisance predictions are generated out-of-sample, to confer a doubly robust

structure (Robins et al., 1994; Chernozhukov et al., 2018). Causal interpretation remains contingent on the identification assumptions: conditional parallel trends, no anticipation, no interference, and adequate overlap.

#### *5.4 Implementation*

For each cohort-event-time comparison, nuisance estimation and forest fitting proceed in two stages. In the first stage, we select the best-performing model for each nuisance function by two-fold cross-validation: candidate models are logistic regression and LASSO logistic regression for the propensity model, and ordinary least squares and elastic-net linear regression for the outcome regression, with two-fold cross-fitting generating the out-of-fold predictions compared across candidates. A causal forest is fit on the full, untrimmed comparison sample using these cross-fitted nuisance predictions; this first-stage forest's propensity estimates are used solely to implement the trimming rule described in Section 5.5, and its treatment-effect estimates are not reported. In the second stage, after trimming, the nuisance model types selected in the first stage are refit on the trimmed sample only, with predictions generated in-sample rather than by a further round of cross-fitting. A causal forest is then re-estimated on the trimmed sample using these in-sample nuisance predictions; this second forest is the one whose treatment-effect estimates, standard errors, and heterogeneity predictions are reported throughout the paper. Consequently, the out-of-sample nuisance-prediction property that motivates cross-fitting, and that the orthogonal moment condition in Section 5.3 is designed to exploit, holds for the nuisance predictions used to decide trimming but not for those used in the forest that produces the reported estimates (Chernozhukov et al., 2018; Athey et al., 2019).

The causal forest itself uses honest sample splitting. In each tree, one subset of the data is used to determine splits in the covariate space, and a separate subset is used to estimate treatment effects within the resulting leaves. This procedure reduces the risk that estimated subgroups are

driven by noise, as the same observations are not used both to define subgroups and to estimate their effects (Athey and Imbens, 2016; Wager and Athey, 2018).

All covariates used for adjustment and heterogeneity estimation are predetermined relative to the cohort-specific baseline period (see 4.2). The forest uses these covariates to learn which baseline patient profiles are associated with larger or smaller DiD effects, thereby enabling HTE to be measured using combinations of pre-treatment characteristics rather than pre-specified subgroup categories.

All analyses were performed using R Statistical Software version 4.2.3 (R: The R Project for Statistical Computing, 2026). The DiDCF estimation was implemented using the *grf* package (Tibshirani et al., 2017).

### *5.5 Heterogeneity summaries, inference, and diagnostics*

For each cohort-event-time comparison, the estimator produces conditional treatment effects  $\hat{\tau}_{c,\ell}(X_i)$  for treated individuals with observed pre-treatment characteristics  $X_i$ . The ATT reported for each cohort-event-time comparison is the forest's doubly robust, augmented inverse-propensity-weighted (AIPW) treated-sample average-treatment-effect estimator, combining the CATT predictions with the fitted propensity and outcome nuisances so that the reported cell-level estimate retains some protection against forest misspecification even where the CATT surface itself is imperfectly estimated. We report average event-time effects, obtained by aggregating these cell-level AIPW ATT estimates across cohorts at each event time  $\ell$ , weighted by each cohort's number of treated observations in that cell, to provide a benchmark for the heterogeneity analyses.

We summarize effect heterogeneity using differences across selected baseline patient characteristics and covariate-binned effect plots. Profile comparisons contrast estimated effects between groups defined by pre-treatment utilization, costs, and prior care histories. Covariate-

binned effect plots complement these comparisons by displaying how  $\hat{\tau}_{c,t}(X_i)$  varies across levels of selected baseline characteristics. Together, these summaries assess whether HTEs are systematically related to observable pre-treatment patient characteristics.

We assess the internal validity of the comparisons using three approaches. First, we report balance diagnostics for key pre-treatment variables. Second, we examine pre-treatment event-study estimates and placebo timing checks to assess whether treated and comparison individuals exhibited similar outcome dynamics prior to enrollment. We test the pretrend and placebo restrictions using a joint Wald test on the pre-treatment and placebo event-time coefficients. Third, we assess overlap and support within each cohort-event-time comparison. In the main analysis, we trim observations with estimated treatment propensities below 0.03 or above 0.97. We retain only comparisons in which at least 60% of observations remain within common support and in which the Kolmogorov-Smirnov distance between the treated and comparison propensity-score distributions does not exceed 0.80. We also require a minimum of 40 treated and 120 comparison observations. Comparisons that do not meet these support requirements are excluded from the corresponding estimand.

We report uncertainty measures that account for repeated observations within individuals. Effects are summarized by event time, with detailed standard errors, confidence intervals, and diagnostic tables provided in Appendix B.

## **6 Results**

We first document selection into IC enrollment and assess the adequacy of support, balance, and pre-treatment diagnostics for the cohort-event-time comparisons. We report average event-time effects to provide a benchmark for subsequent analyses. The main analysis evaluates whether these averages mask systematic heterogeneity in treatment effects by baseline patient characteristics. Finally, we summarize the results of robustness and sensitivity analyses.

## 6.1 Selection into enrollment and empirical support

Baseline characteristics indicate substantial observable selection into IC enrollment. Enrolled individuals differ from non-enrolled individuals in demographic characteristics, pre-enrollment care use, and costs. They are more frequently female and have higher pre-enrollment utilization and costs, including hospital admissions, ambulatory visits, inpatient costs, medication costs, and total healthcare costs. Table 1 presents the corresponding baseline means and standardized mean differences. These differences motivate the use of pre-treatment adjustment, baseline-differenced outcomes, and common-support restrictions within each cohort-event-time comparison.

Figure A1 displays covariate balance before and after propensity-score adjustment. At the enrollment quarter  $\ell = 0$ , the mean absolute standardized mean difference across covariates decreases from 0.121–0.132 before adjustment to 0.023–0.038 after adjustment, depending on the outcome. This indicates substantially improved balance in observed pre-treatment characteristics.

**Table 1** Baseline characteristics by enrollment status

| Variable               | Enrolled             | Not enrolled         | Total sample         | SMD   |
|------------------------|----------------------|----------------------|----------------------|-------|
| N individuals          | 4,336                | 45,851               | 50,187               |       |
| Age (years)            | 47.54 (20.94)        | 48.66 (23.79)        | 48.56 (23.56)        | -0.05 |
| Female                 | 61.9%                | 49.8%                | 50.9%                | 0.25  |
| Employed               | 27.0%                | 34.1%                | 33.5%                | -0.15 |
| Retired                | 22.4%                | 28.0%                | 27.5%                | -0.13 |
| Hospital admissions    | 0.16 (0.32)          | 0.09 (0.24)          | 0.10 (0.25)          | 0.24  |
| Any ACSC admission     | 0.1%                 | 0.1%                 | 0.1%                 | 0.00  |
| Outpatient visits      | 3.31 (2.23)          | 1.94 (1.79)          | 2.06 (1.87)          | 0.68  |
| Outpatient costs       | 285.17 (420.72)      | 161.82 (369.92)      | 172.26 (376.05)      | 0.31  |
| Inpatient costs        | 187.52<br>(1,409.94) | 60.75 (581.93)       | 71.49 (692.47)       | 0.12  |
| Medication costs       | 248.12<br>(1,856.84) | 165.47<br>(1,256.48) | 172.47<br>(1,318.13) | 0.05  |
| Total healthcare costs | 937.07<br>(2,748.52) | 706.01<br>(1,923.25) | 725.57<br>(2,007.33) | 0.10  |



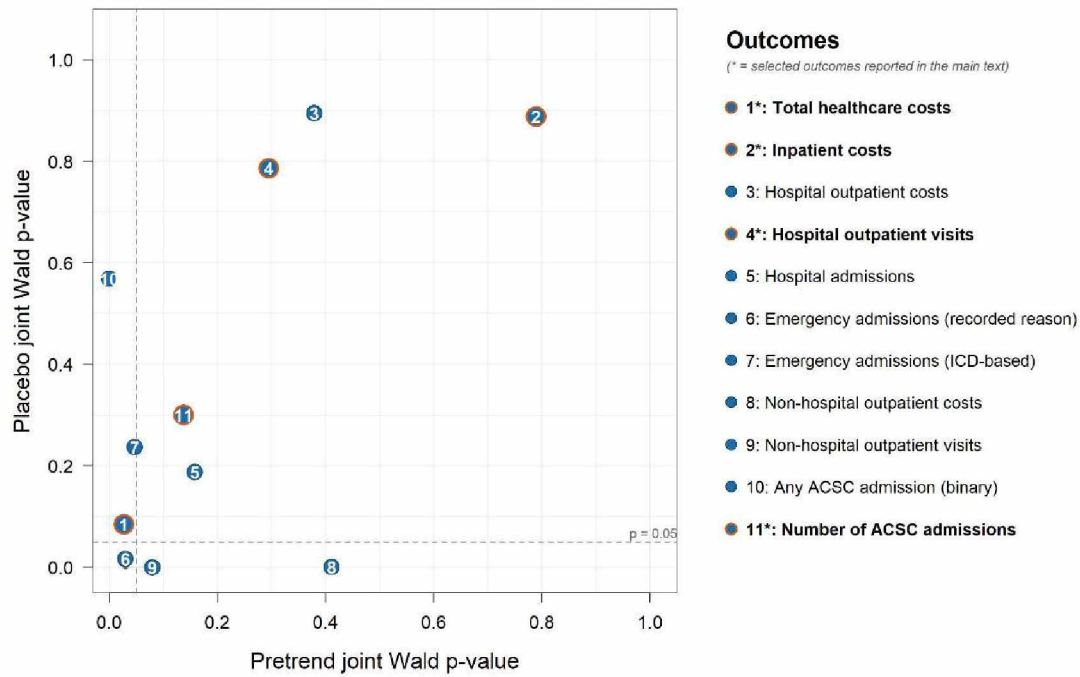
Notes: The table reports selected characteristics by enrollment status in the target-district sample. Continuous variables are mean (standard deviation); binary variables are percentages. Utilization, count, cost, and ACSC variables are measured over the pre-enrollment baseline window. Utilization and count variables are means per quarter; cost variables are euros per quarter. Any ACSC admission indicates the share of individuals with at least one ambulatory care sensitive condition admission during the pre-enrollment baseline window. SMD denotes the standardized mean difference, enrolled minus not enrolled. Outpatient refers to the non-hospital setting.

The staggered design requires comparable untreated observations for each enrollment cohort and event time. Not every observed enrollment cohort contributes to every event time, and some cohorts are excluded from the final estimated effects after support restrictions are applied, as estimation requires sufficiently large and comparable untreated groups within each cohort-event-time comparison. In the main analysis, 23 cohorts contribute to at least one estimated comparison. After propensity-score and common-support restrictions, the estimated cohort-event-time cells contain 40 to 416 treated observations and 230 to 4,820 comparison observations across the main outcome specifications.

Support diagnostics indicate that identification is partial and cell-specific, consistent with the staggered observational design. For each outcome, there are 377 possible cohort-event-time comparisons before support filters, based on 29 enrollment cohorts and 13 event times. After applying the pre-specified restrictions described in Section 5, the number of retained comparisons ranges from 239 to 250 across outcomes, corresponding to retained-cell shares of approximately 0.63 to 0.66. Average common-support shares in retained cells are high, ranging from 94.2% to 95.0%, and mean treated-comparison Kolmogorov-Smirnov distances remain below the pre-specified threshold of 0.80.

Pretrend and placebo diagnostics are outcome-specific and are assessed using joint Wald tests on the pre-treatment event-time coefficients and on the placebo timing coefficients. Figure 1 reports the resulting joint-test p-values by outcome. Inpatient costs, hospital outpatient costs, hospital outpatient visits, and ACSC admissions show no evidence against parallel pretrends and no evidence against the placebo timing restriction. Hospital admissions likewise show no

evidence against either diagnostic. Total healthcare costs present an asymmetric picture: the joint pretrend test rejects ( $p=0.027$ ), while the placebo test does not. Non-hospital ambulatory visits fail both diagnostics. ICD-based emergency admissions, any ACSC admission, and record-based emergency admissions each fail the pre-trend test but not the placebo test. Non-hospital ambulatory costs pass the pre-trend test but fail the placebo test. Record-based emergency admissions fail both the pre-trend and placebo tests. Table A3 reports the full joint-test statistics by outcome.



**Figure 1** Pre-treatment trends and placebo diagnostics by outcome Notes: Each point is one outcome. The x-axis is the joint Wald test p-value for the pretrend restriction (chi-square, 3 degrees of freedom, event times -4/-3/-2 relative to enrollment, B=399 cohort-resampling bootstrap replicates). The y-axis is the joint Wald test p-value for the placebo restriction (chi-square, 2 degrees of freedom, event times -3/-2 under a placebo treatment date shifted 2 quarters earlier; the third placebo event time, -1, is the omitted reference period). Dashed lines mark  $p=0.05$  on each axis for visual reference only. Point border color and size distinguish the four outcomes reported as selected results in the main analysis.

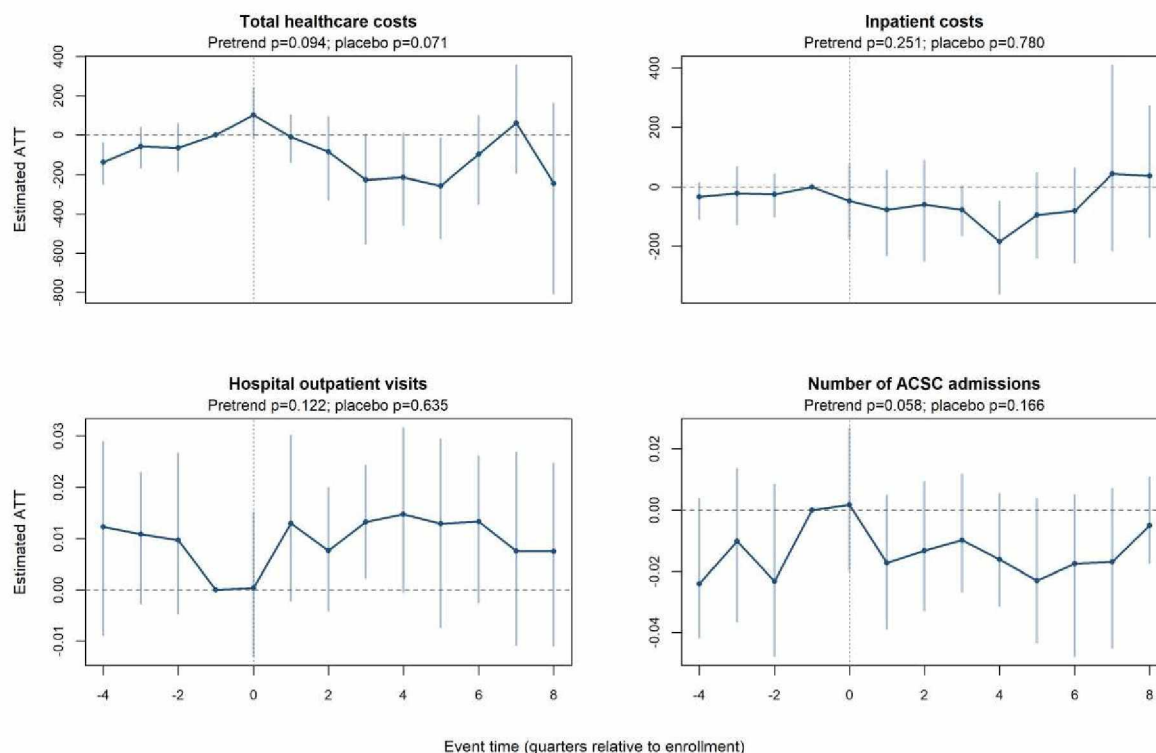
## 6.2 Average event-time effects

Figure 2 presents average event-time effects for selected outcomes from four quarters before to eight quarters after enrollment; Figure A3 provides results for all outcomes. These estimates

serve as the reference for subsequent heterogeneity analyses. Average effects are generally modest and mixed. Total healthcare costs remain close to zero with wide confidence intervals and no consistent post-enrollment reduction. Inpatient cost estimates are predominantly negative in the early and middle post-enrollment periods, though imprecise. Hospital outpatient costs remain near zero, while hospital outpatient visits display small positive post-enrollment effects. Based on pointwise 95% bootstrap confidence intervals, only a few estimates, total costs at  $\ell = 5$ , inpatient costs at  $\ell = 4$ , and hospital outpatient visits at  $\ell = 3$ , exclude zero. None of these results indicate a sustained post-enrollment trend.

In summary, the average event-time effects do not show a consistent reduction in hospital-based utilization or costs following enrollment. Average changes in inpatient and cost-related outcomes are limited, while increases are more apparent for some ambulatory outcomes. The

following sections assess whether these averages conceal systematic variation across baseline patient characteristics.



**Figure 2** Event-time average treatment effects for selected outcomes

Notes: Points are treated-count-weighted average treatment effects across estimable cohort–event-time comparisons. Event time is measured in quarters relative to enrollment:  $-1$  is the baseline/reference quarter,  $0$  is the enrollment quarter, and positive values are post-enrollment quarters. Vertical bars are 95% intervals from 399 cohort-resampling bootstrap replications. Panels use outcome-specific vertical scales. Selected outcomes are total healthcare costs, inpatient costs, hospital outpatient visits, and the number of ambulatory care-sensitive condition (ACSC) admissions; each passed the pretrend and placebo diagnostic thresholds.

### 6.3 Distribution of treatment effects

Figure A4 displays the distribution of estimated CATTs among enrolled individuals for selected outcomes. Figure A5 reports the corresponding standardized distributions for all eleven outcomes on a common scale. The distributions show substantial dispersion, particularly for cost and inpatient-related outcomes.

For total healthcare costs, estimated CATTs are widely dispersed, with a weighted mean of approximately -122 euros and a weighted median of approximately -89 euros among treated observations in the post-period summary. The distribution includes both large negative and large positive estimated effects, indicating substantial heterogeneity in predicted cost responses across patient profiles. Inpatient costs exhibit substantial mass near zero but also long tails, again suggesting that some enrolled individuals experience much larger estimated reductions or increases than the average treatment effect would imply.

For utilization outcomes such as hospital admissions and emergency admissions, the absolute scale is smaller, but the distributions remain dispersed around zero. These findings indicate that heterogeneity in treatment effects is not limited to cost outcomes, although the most pronounced dispersion is observed in inpatient and cost-related outcomes.

Figure A6 plots the CATT by predicted-effect quartiles, ranking treated patients by their predicted effects  $\tau_{c,\ell}(x)$  and summarizing the distribution of estimated treatment effects at each event time. The largest contrasts are observed in inpatient and cost-related outcomes. At event time  $\ell = 1$ , the first quartile shows decreases of 560 euros in total costs and 593 euros in inpatient costs, while the fourth quartile shows increases of 544 euros and 364 euros,

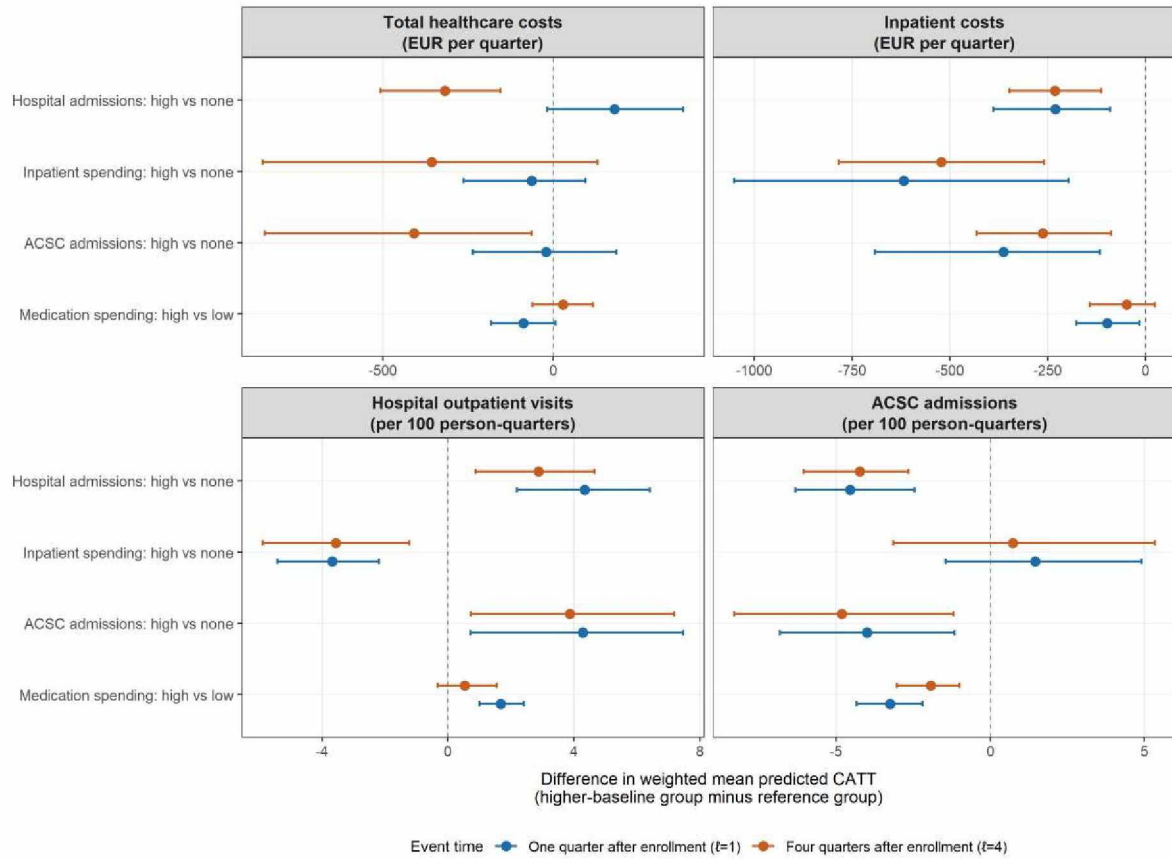
respectively. Similar separations are evident for hospital admissions and ACSC outcomes. Figure A7 reports the corresponding quartile paths for all eleven outcomes.

These distributional results indicate that average treatment effects summarize only part of the estimated effect distribution and do not identify which participants experience stronger or weaker effects. The next section relates estimated CATTs to baseline utilization, costs, and care-contact histories to identify associated patient characteristics.

#### *6.4 Heterogeneity across baseline patient characteristics*

Figures 3 and 4 provide complementary evidence on which observable patient profiles are associated with larger or smaller estimated effects. Figure 3 compares the highest and lowest baseline categories at one and four quarters after enrollment. For prior admissions and inpatient spending, the reference group comprises patients with no recorded baseline use or costs; for medication costs, the reference is the lowest quartile.

Patients with at least three prior hospital admissions had inpatient costs approximately 230 euros lower than those without prior admissions at both event times. Their estimated reduction in ACSC admissions was also about four admissions greater per 100 person-quarters. Prior ACSC admissions showed a similar pattern. Patients in the highest baseline inpatient-cost group experienced particularly large reductions in inpatient costs, although these estimates were less precise due to the group's small size. Differences in total costs became more pronounced four quarters after enrollment, especially among patients with substantial prior hospital or ACSC use.



**Figure 3** Differences in Estimated Treatment Effects across Baseline Patient Profiles

Notes: Points show differences in weighted mean predicted CATTs among treated patients (higher-baseline minus reference group) one and four quarters after enrollment; enrollment is event time 0. Negative cost and ACSC contrasts indicate larger reductions in the higher-baseline group; positive outpatient contrasts indicate larger increases. Baseline measures cover four pre-enrollment quarters. Higher admission use means at least three hospital or ACSC admissions versus none. Positive inpatient cost was divided into empirical tertiles; higher cost means above the upper tertile cutoff of EUR 1,181 per quarter, versus zero. Medication cost comparison between Q4 (above EUR 237) and Q1 (up to EUR 17). Cut-points are fixed across outcomes, and ties remain in the same group. Horizontal bars are pointwise 95% intervals from 300 individual-cluster bootstrap replications, conditional on the fitted forest. ACSC denotes ambulatory care-sensitive conditions.

Figure 4 examines broader gradients in estimated effects across baseline covariate levels over the post-enrollment period. CATT predictions are pooled across event times  $\ell = 1, \dots, 8$ ; each point represents the weighted mean predicted CATT for treated observations within a covariate bin across these post-enrollment quarters, rather than an effect for a single quarter. Cost effects generally became more negative as prior IC contact increased. Higher baseline outpatient use was also associated with larger reductions in total and inpatient costs, although this association

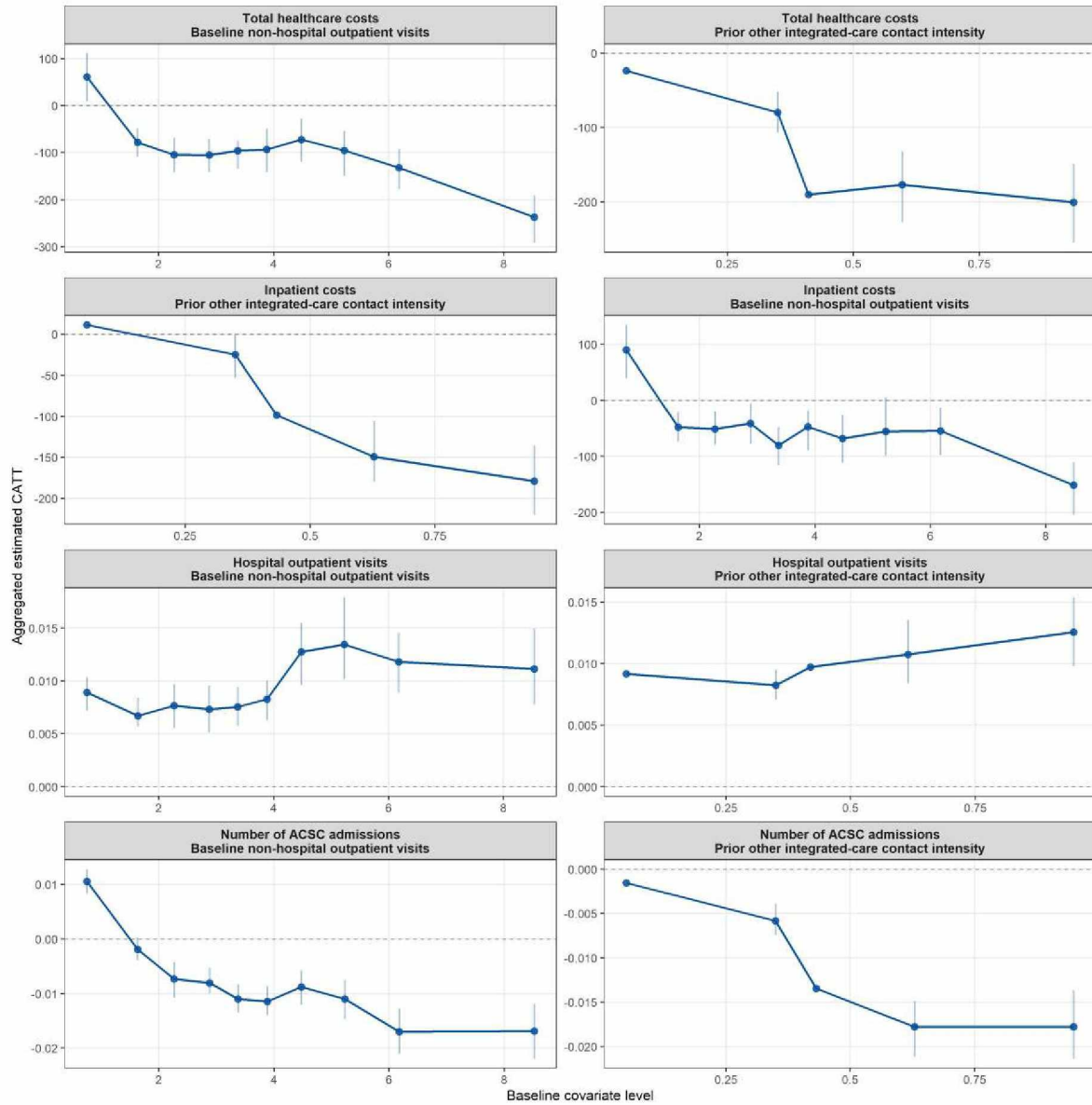
was less consistent. For ACSC admissions, both greater baseline outpatient use and more intensive prior IC contact were associated with larger estimated reductions.

Hospital outpatient visits display a different pattern. Estimated increases were generally larger among patients with greater baseline outpatient use or more intensive prior IC contact. These results indicate that lower inpatient and ACSC utilization may coincide with increased outpatient contact for some patient profiles, rather than an overall reduction in healthcare use.

Taken together, Figures 3 and 4 indicate that the largest estimated decreases are concentrated among patients with substantial prior acute-care use and broader engagement with healthcare and integrated-care services. No single baseline measure consistently identifies larger decreases across all outcomes.

Table B1 shows that, across several outcomes, more favorable estimated effects are associated with higher baseline non-hospital ambulatory costs, higher medication costs, greater prior utilization, or prior IC contact.





**Figure 4** Estimated HTEs by baseline covariate level

Notes: Points are weighted mean predicted CATTs among treated observations, pooled across post-enrollment event times  $\ell = 1, \dots, 8$ , within empirical covariate bins. Horizontal positions show mean covariate values within bins. Unless otherwise stated, baseline utilization and spending measures are averages per quarter over the four quarters preceding enrollment. Prior other integrated-care contact intensity is the share of all observed pre-enrollment quarters with participation in another integrated-care contract. For each outcome, the figure shows the two highly ranked covariates with the largest range in binned mean CATTs. Bars are pointwise 95% individual-cluster bootstrap intervals where available, conditional on the fitted forest. Baseline outpatient visits are four-quarter pre-enrollment means. ACSC denotes ambulatory care-sensitive conditions.

Additional heterogeneity summaries are reported in Appendix B. Best linear projection estimates (Tables B2-B5) and variable-importance rankings (Figure B1) broadly support these

results, consistently identifying prior utilization and costs, ACSC history, and prior participation in other IC programs as key dimensions of heterogeneity in estimated treatment effects. Targeting Operator Characteristic (TOC) curves (Figures B2 and B3) assess whether ranking by predicted CATT identifies treated observations with effects that differ from the average. Figure B4 reports the corresponding covariate-level gradients of Figure 4 for all eleven outcomes.

### *6.5 Robustness and sensitivity*

We assess robustness using several prespecified dimensions. The analyses vary treatment timing and anticipation assumptions by implementing alternative anticipation windows and placebo timing. We apply alternative sample definitions, including the full sample, strictly balanced panels, adult-only restrictions, exclusion of individuals enrolled while in hospital, and exclusions of the acute COVID-19 period. Additional analyses consider alternative treatment definitions based on formal enrollment, consultation, and provider-network contact. We also vary common-support criteria and inference procedures, including stricter and looser support thresholds, provider-level clustering, and bootstrap inference. Finally, we examine alternative outcome constructions for cost, ACSC admission, emergency admission, and ambulatory utilization outcomes, including adjustment for price changes.

The robustness checks are consistent with the main qualitative findings. Average event-time effects remain modest. HTE analyses continue to indicate larger reductions in inpatient costs and ACSC admissions among patients with higher baseline prior hospital or ACSC use and inpatient costs. However, the robustness checks do not eliminate the outcome-specific diagnostic differences. Results for ambulatory outcomes and certain binary outcomes should therefore be interpreted with greater caution. Detailed robustness and sensitivity analyses are presented in Appendix C.

## **7 Discussion**

### *7.1 Main findings and economic interpretation*

The primary economic finding is that average program effects do not fully capture the value of IC. In the evaluated initiative in a socially deprived urban area in Germany, average treatment effects on inpatient utilization and cost-related outcomes are mixed and generally modest. Ambulatory outcomes indicate more consistent increases, although these estimates are supported by weaker pre-trend and placebo diagnostics. Total and inpatient costs likewise differ in diagnostic strength: inpatient costs shows no evidence against either diagnostic, while total costs shows evidence against parallel pretrends specifically. Estimated effects differ substantially across supported patient profiles. The most pronounced heterogeneity in treatment effects is observed for total and inpatient costs, with larger contrasts between predicted-effect quartiles for hospital admissions and ACSC admissions. Estimated decreases are concentrated among individuals with higher pre-period care intensity, higher prior costs, ACSC history, and previous contact with other care programs. In contrast, profiles with lower baseline utilization and costs exhibit weaker or negligible responses.

This distribution is relevant because the economic interpretation of IC depends on how utilization and cost effects are allocated across patient groups. A modest average treatment effect may reflect reductions in hospital-based care for some individuals alongside limited effects, increased outpatient use, or higher costs for others. For resource-intensive care coordination and navigation services, average effects alone provide limited information for assessing program value, scalability, and resource allocation.

The findings also indicate a tension between cost-focused interpretation and the broader equity objectives of IC in socially deprived areas. From a claims-based cost perspective, patients with higher baseline morbidity, prior acute-care use, and higher costs may generate the largest measurable reductions in hospital-based care. However, IC programs may also seek to improve

access, navigation, patient experience, and trust among groups whose short-run claims-based cost effects are limited. Patients with lower baseline utilization should not be classified as low-value beneficiaries solely on the basis of claims-based outcomes. The results suggest that claims-based economic value and broader access-oriented value are not necessarily aligned.

### *7.2 Relation to previous evidence*

These findings contribute to previous national and international evaluations. Prior studies in this setting indicate that the initiative can alter service use and healthcare utilization, including changes in primary-care demand and increases in several utilization and cost outcomes. The present results further show that these changes are not uniformly distributed across enrolled individuals. Estimated responses vary according to pre-treatment care histories, prior costs, ACSC history, and other indicators of care need.

IC interventions may increase certain types of utilization by improving access, identifying unmet needs, or intensifying monitoring, while reducing costly or avoidable care for selected patient profiles. These mechanisms can operate concurrently. Consequently, average effects may appear small, mixed, or cost-increasing, even when some patient profiles exhibit estimated reductions in hospital-based care or costs. HTE estimation is therefore informative not only for describing subgroup variation, but also for explaining why average effect evaluations of complex care programs often yield mixed results.

### *7.3 Limitations*

Several limitations should be considered. Identification relies on conditional parallel trends in untreated outcome changes. Although the analysis adjusts for pre-treatment claims histories and assesses pretrend, placebo, overlap, and balance diagnostics, unobserved time-varying factors may still affect enrollment and outcomes. Support restrictions also imply that the estimated effects apply to cohort-event-time comparisons with adequate overlap between enrolled and

untreated observations. The results describe effects within the supported comparisons and should not be generalized to all enrolled individuals or to any possible enrollment cohort.

The analysis relies on administrative claims data and therefore measures billed utilization and costs, but does not capture all dimensions of healthcare quality, patient experience, or unmet need. Some increases in utilization may reflect improved access rather than inefficient use. This is particularly relevant for ambulatory outcomes, where estimated increases may indicate care activation or improved access rather than overuse.

The heterogeneity summaries describe associations between baseline patient profiles and estimated treatment effects and should not be interpreted as causal effects of the baseline covariates themselves. The analysis estimates HTEs among enrolled individuals and provides evidence relevant to subgroup-sensitive program design. However, it does not determine an optimal allocation rule for future patients.

#### *7.4 Conclusion*

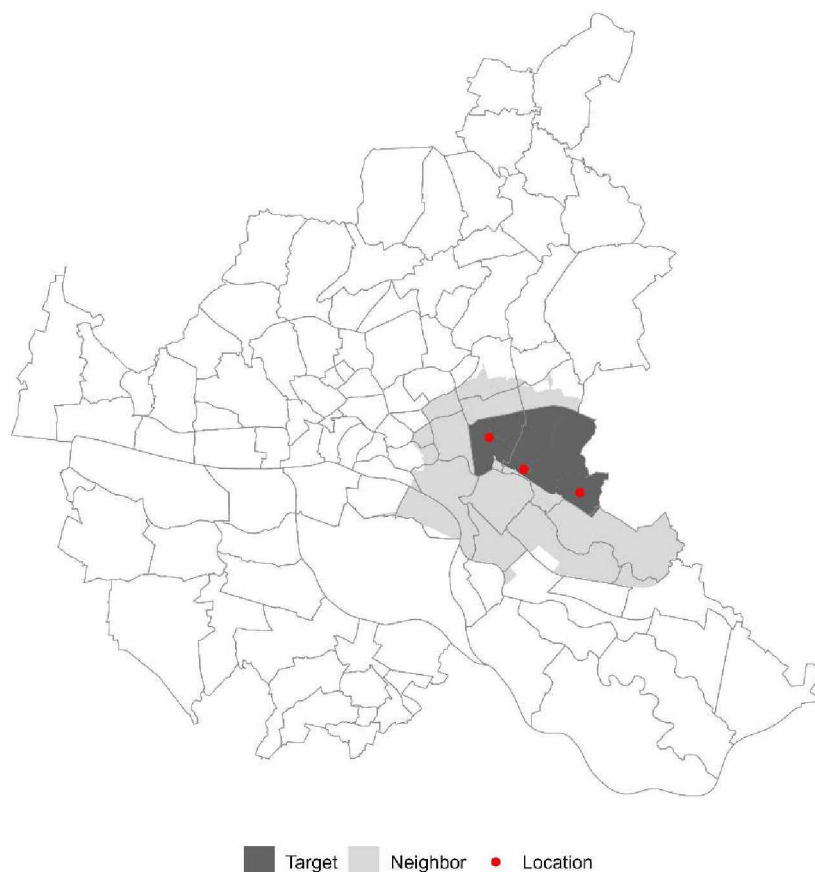
This study demonstrates that economic evaluation of IC is altered when heterogeneity in treatment effects is explicitly considered. In the evaluated program, average effects are mixed and generally modest, but estimated effects vary substantially across patient profiles. The most pronounced heterogeneity is observed in total and inpatient costs, with larger contrasts for hospital admissions and ACSC admissions. Estimated decreases are concentrated among patients with higher pre-period care intensity, higher prior costs, ACSC history, and previous contact with other IC programs.

These findings indicate that average effects alone are insufficient for assessing the value of IC. For health-economic evaluation, it is important to identify which patient groups generate changes in hospital-based care, outpatient use, avoidable admissions, and costs. Credible HTE estimation can provide important evidence for interpreting program value. The estimated

heterogeneity also offers empirical input for future work on subgroup-sensitive program design and policy-learning analyses, while any allocation rule would require further assumptions and explicit consideration of capacity, feasibility, and equity.

## Appendix A. Outcomes, identification, and HTE support

This appendix defines the outcomes and covariates, then reports which cohort-event-time comparisons are estimable, presents average event-time effects as benchmarks, and finally summarizes the distribution and evolution of estimated heterogeneous treatment effects.



**Figure A1** Intervention area and neighboring districts in Hamburg

### A.1 Outcomes and covariates

Table A1 defines the quarterly utilization and spending outcomes. Table A2 then describes the pre-treatment characteristics used for adjustment and for estimating heterogeneity. The covariates are predetermined relative to each cohort's baseline quarter; most utilization and cost histories are four-quarter rolling pre-treatment means.

**Table A1** Outcomes used in the DiD-causal-forest analysis

| Outcome                         | Definition                                                                                                        |
|---------------------------------|-------------------------------------------------------------------------------------------------------------------|
| Hospital outpatient costs       | Quarterly hospital outpatient costs                                                                               |
| Non-hospital ambulatory costs   | Quarterly non-hospital ambulatory costs                                                                           |
| Total healthcare costs          | Quarterly total healthcare costs                                                                                  |
| Emergency admissions, ICD-based | Quarterly emergency admission measure based on ICD-coding, according to Krämer et al. (2019)(Krämer et al., 2019) |

|                                       |                                                                                           |
|---------------------------------------|-------------------------------------------------------------------------------------------|
| Emergency admissions, recorded reason | Quarterly emergency admission measure based on recorded admission reason                  |
| Any ACSC admission                    | Quarterly binary indicator for at least one ambulatory care sensitive condition admission |
| Number of ACSC admissions             | Quarterly count of ambulatory care sensitive condition admissions                         |
| Hospital admissions                   | Quarterly hospital admission count                                                        |
| Hospital outpatient visits            | Quarterly hospital outpatient visit count                                                 |
| Non-hospital ambulatory visits        | Quarterly non-hospital ambulatory visit count                                             |

Notes: ACSC denotes ambulatory care sensitive conditions. ICD denotes International Classification of Diseases. DiD denotes difference-in-differences. Outcomes are measured quarterly. In the DiD-causal-forest analysis, effects are estimated for changes relative to the cohort-specific baseline quarter. Absence of sector-specific utilization is coded as zero rather than treated as missing.

**Table A2** Baseline covariates used in the DiD-causal-forest analysis

| Covariate                      | Definition                                                                                             |
|--------------------------------|--------------------------------------------------------------------------------------------------------|
| Long-term care costs           | Long-term care costs measured as a four-quarter rolling mean ending in the baseline quarter.           |
| Age                            | Measured in the cohort-specific baseline quarter.                                                      |
| Female                         | Binary sex indicator measured before treatment.                                                        |
| Medication costs               | Medication costs measured as a four-quarter rolling mean ending in the baseline quarter.               |
| Any ACSC admission             | Indicator or mean for any ACSC admission in the four-quarter baseline window.                          |
| ACSC admissions                | Number of ACSC admissions measured as a four-quarter rolling mean ending in the baseline quarter.      |
| Prior DMP participation        | Share of observed pre-treatment quarters with DMP participation.                                       |
| Prior IC participation         | Share of observed pre-treatment quarters with participation in other IC contracts.                     |
| Hospital admissions            | Hospital admissions measured as a four-quarter rolling mean ending in the baseline quarter.            |
| Hospital outpatient costs      | Hospital outpatient costs measured as a four-quarter rolling mean ending in the baseline quarter.      |
| Inpatient costs                | Inpatient costs measured as a four-quarter rolling mean ending in the baseline quarter.                |
| Non-hospital ambulatory costs  | Non-hospital ambulatory costs measured as a four-quarter rolling mean ending in the baseline quarter.  |
| Non-hospital ambulatory visits | Non-hospital ambulatory visits measured as a four-quarter rolling mean ending in the baseline quarter. |

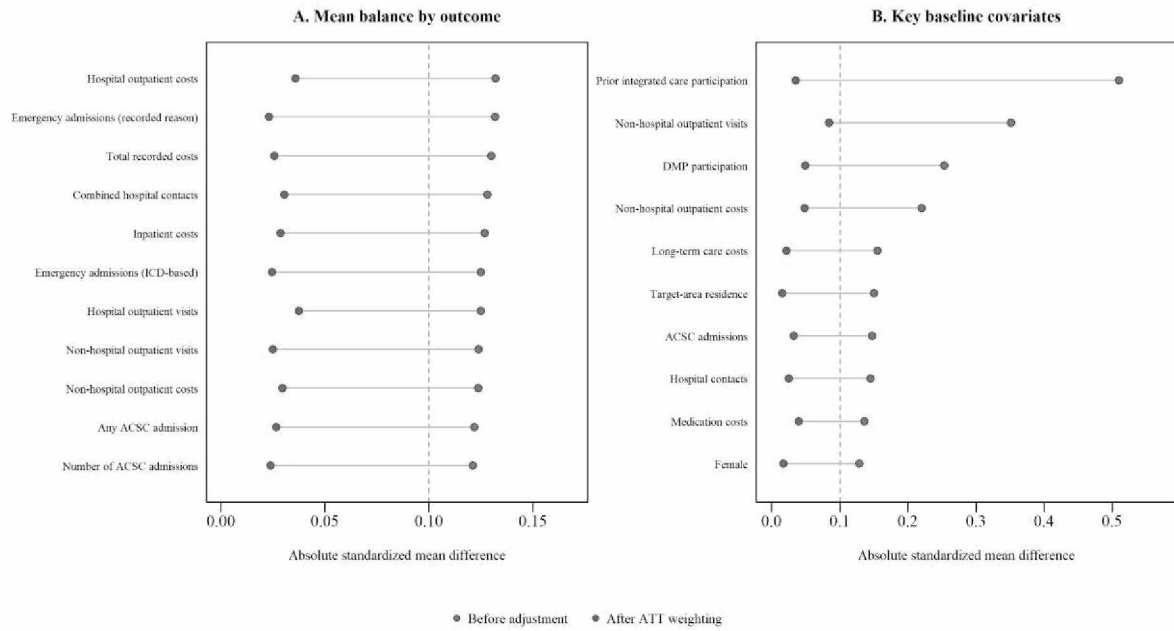


|                        |                                                                                                                                       |
|------------------------|---------------------------------------------------------------------------------------------------------------------------------------|
| Sickness-absence costs | Sickness-absence costs measured as a four-quarter rolling mean ending in the baseline quarter, defined for at-risk employed quarters. |
| Sickness-absence days  | Sickness-absence days measured as a four-quarter rolling mean ending in the baseline quarter, defined for at-risk employed quarters.  |
| Residence              | Indicator for residence in the target area used in the analysis data.                                                                 |

Notes: The table lists the 16 covariates included in the main DiD causal forest. All covariates are predetermined relative to the cohort-specific baseline quarter, defined as the quarter immediately before enrollment; comparison individuals are measured in the same calendar periods. Unless otherwise stated, utilization, cost, ACSC, and sickness-absence measures are four-quarter means ending in the baseline quarter. Costs are euros per quarter. Any ACSC admission indicates at least one admission during this four-quarter window. Age is measured in the baseline quarter, sex is time-invariant, and target-area residence is measured before enrollment. Prior disease-management and other integrated-care participation are shares across all observed pre-enrollment quarters. ACSC denotes ambulatory care-sensitive conditions; DMP denotes disease-management program. Absence of sector-specific utilization is coded as zero.

## A.2 Covariate balance

Figure A1 compares covariate balance before and after propensity-score adjustment in the enrollment quarter  $\ell = 0$ . Across outcomes, the treated-count-weighted mean absolute standardized mean difference falls from approximately 0.12-0.13 before adjustment to 0.02-0.04 afterward. The largest initial differences occur for prior participation in other integrated-care contracts, non-hospital outpatient use, disease-management program participation, and outpatient costs; each is substantially reduced after adjustment.



**Figure A2** Covariate balance before and after adjustment

Notes: Panel A reports the treated-count-weighted mean absolute standardized mean difference across covariates and estimable cells at  $\ell = 0$ . Panel B shows the ten substantive covariates with the largest initial imbalance in the total-cost analysis; missingness indicators are omitted. Before adjustment, observations receive equal weight. After adjustment, controls receive estimated propensity-odds weights  $\hat{p}(X)/(1 - \hat{p}(X))$  and treated observations unit weight. The dashed line marks 0.10. Weighting is a balance diagnostic, not a separate estimator; effects are estimated by the DiD causal forest. ACSC denotes ambulatory care-sensitive conditions; DMP denotes disease-management program. Prior integrated care and DMP participation are measured as shares of observed pre-enrollment quarters with recorded participation.

## A.3 Identification and diagnostic support

Table A3 reports candidate and retained cohort-event-time comparisons, average overlap, and the weakest pretrend and placebo diagnostics for each outcome. The table should be read before the effect figures: outcomes with failed diagnostics receive correspondingly more cautious interpretation.

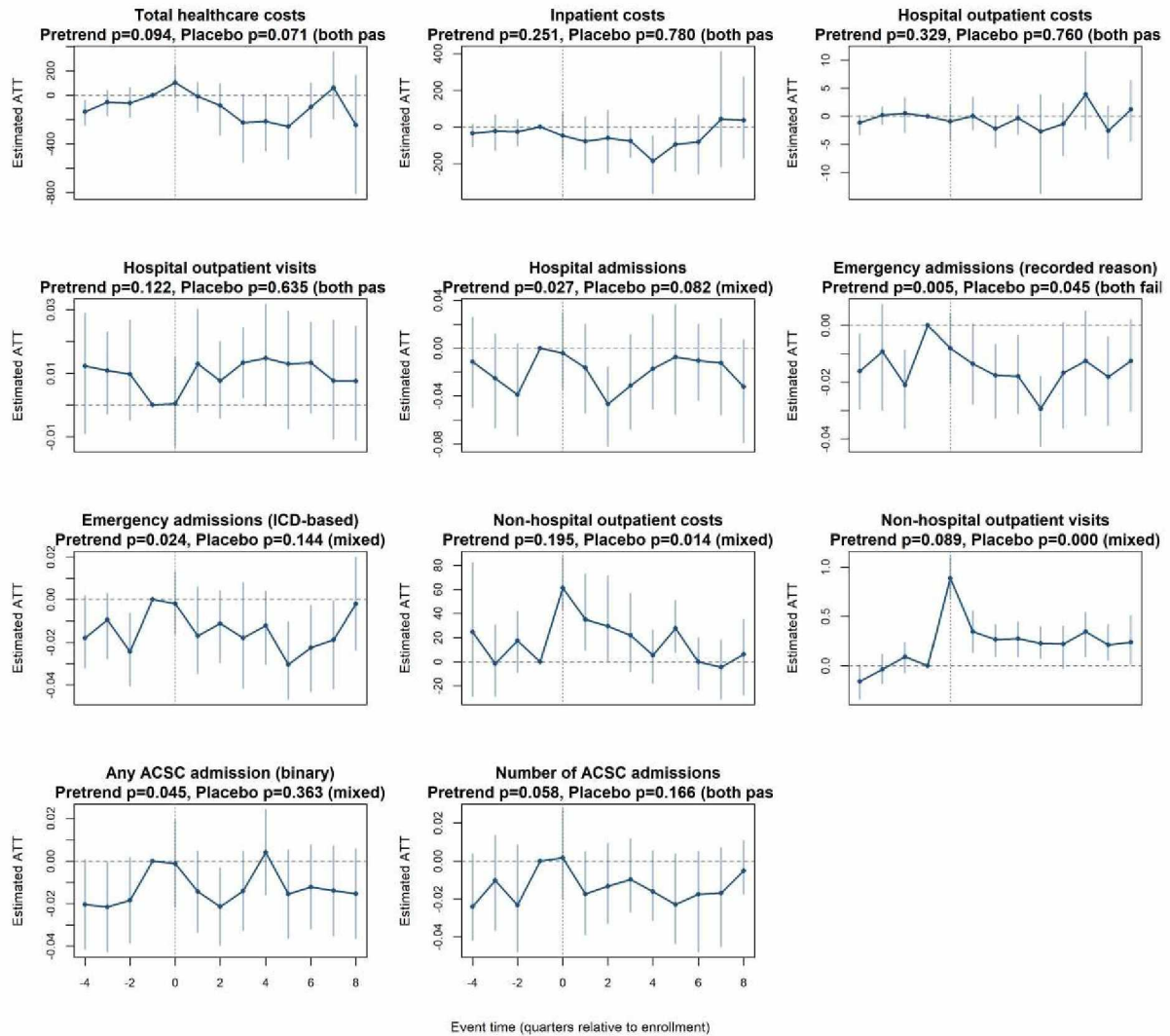
**Table A3** Identification Summary by Outcome

| Outcome # | Obs. cohorts | Overlap share | KS    | Pretrend joint Wald p | Placebo joint Wald p |
|-----------|--------------|---------------|-------|-----------------------|----------------------|
| 1         | 29           | 94.9%         | 0.452 | 0.027                 | 0.085                |
| 2         | 29           | 94.8%         | 0.452 | 0.789                 | 0.888                |
| 3         | 29           | 95.0%         | 0.451 | 0.379                 | 0.895                |
| 4         | 29           | 94.9%         | 0.449 | 0.295                 | 0.787                |
| 5         | 29           | 94.8%         | 0.448 | 0.158                 | 0.187                |
| 6         | 29           | 94.4%         | 0.448 | 0.030                 | 0.017                |
| 7         | 29           | 94.7%         | 0.452 | 0.047                 | 0.237                |
| 8         | 29           | 94.2%         | 0.452 | 0.411                 | 0.000                |
| 9         | 29           | 94.8%         | 0.454 | 0.030                 | 0.000                |
| 10        | 29           | 94.2%         | 0.451 | 0.000                 | 0.569                |
| 11        | 29           | 94.5%         | 0.456 | 0.137                 | 0.301                |

Notes: Outcomes are numbered as follows: 1 = total healthcare costs, 2 = inpatient costs, 3 = hospital outpatient costs, 4 = hospital outpatient visits, 5 = hospital admissions, 6 = emergency admission (record-based), 7 = emergency admission (ICD-based), 8 = Non-hospital ambulatory costs, 9 = Non-hospital ambulatory visits, 10 = any ambulatory care sensitive condition (ACSC) admission, 11 = ACSC admissions. Cells are cohort-event-time comparisons. Estimation requires at least 40 treated and 120 comparison observations, propensity-score support between 0.03 and 0.97, at least 60% of observations remaining in common support, and a treated-comparison Kolmogorov-Smirnov (KS) distance no larger than 0.80. Pretrend and placebo columns report the worst supported diagnostic p-value by outcome; values below 0.05 indicate weaker diagnostic support.

## A.4 Average event-time effects across outcomes

Figure A3 provides the complete set of average event-time benchmarks. These estimates summarize average changes among treated patients and do not describe how effects vary across patient profiles. The diagnostic subtitles make clear why the main text places greater weight on outcomes with stronger pretrend and placebo support.

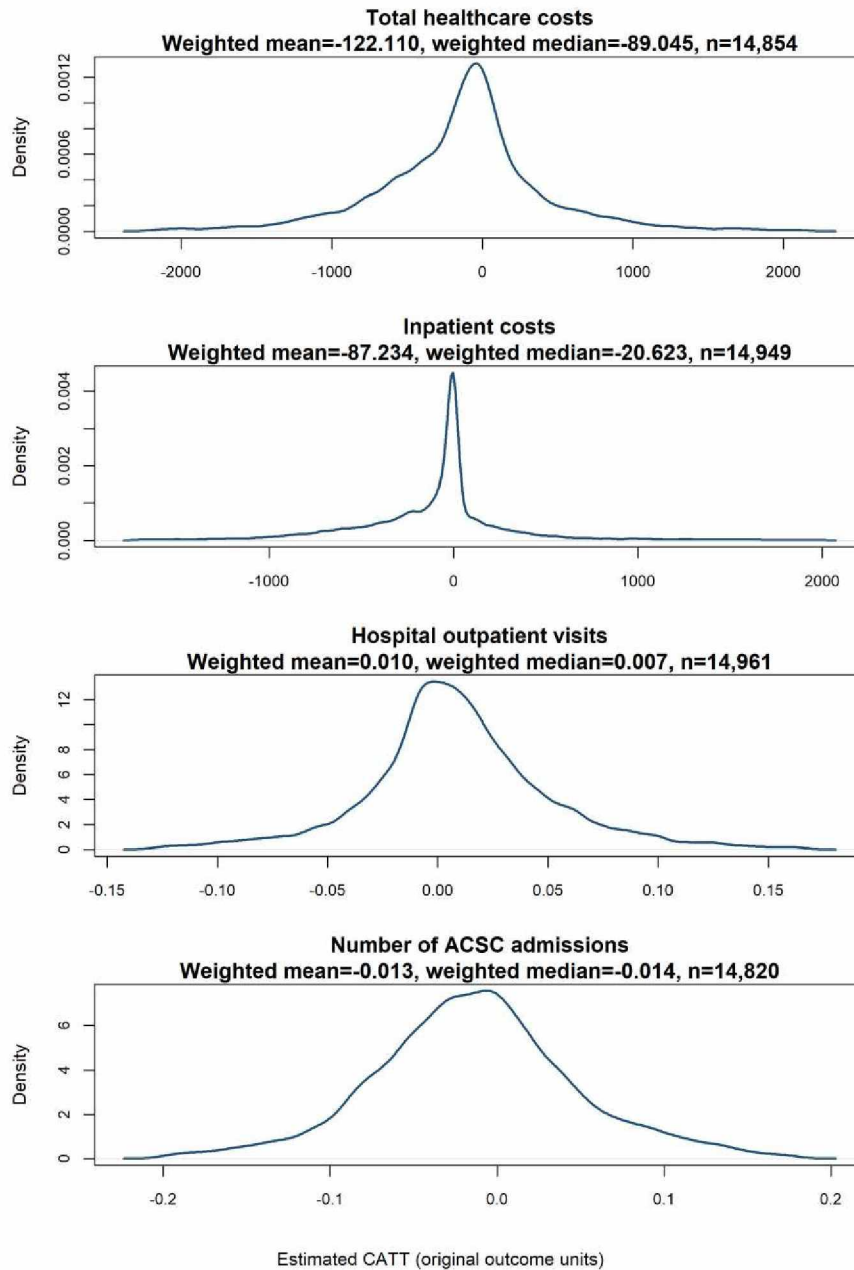


**Figure A3** Average treatment effects over event time for outcomes

Notes: Points are treated-count-weighted average treatment effects across estimable cohort-event-time comparisons; vertical bars are 95% intervals from 399 cohort-resampling bootstrap replications. Event time is measured in quarters relative to enrollment:  $\ell = -1$  is the baseline quarter,  $\ell = 0$  is the enrollment quarter, and positive values are post-enrollment quarters. Panels use outcome-specific vertical scales. Subtitles report the worst supported pretrend and placebo p-values. ACSC denotes ambulatory care-sensitive conditions; ICD denotes International Classification of Diseases.

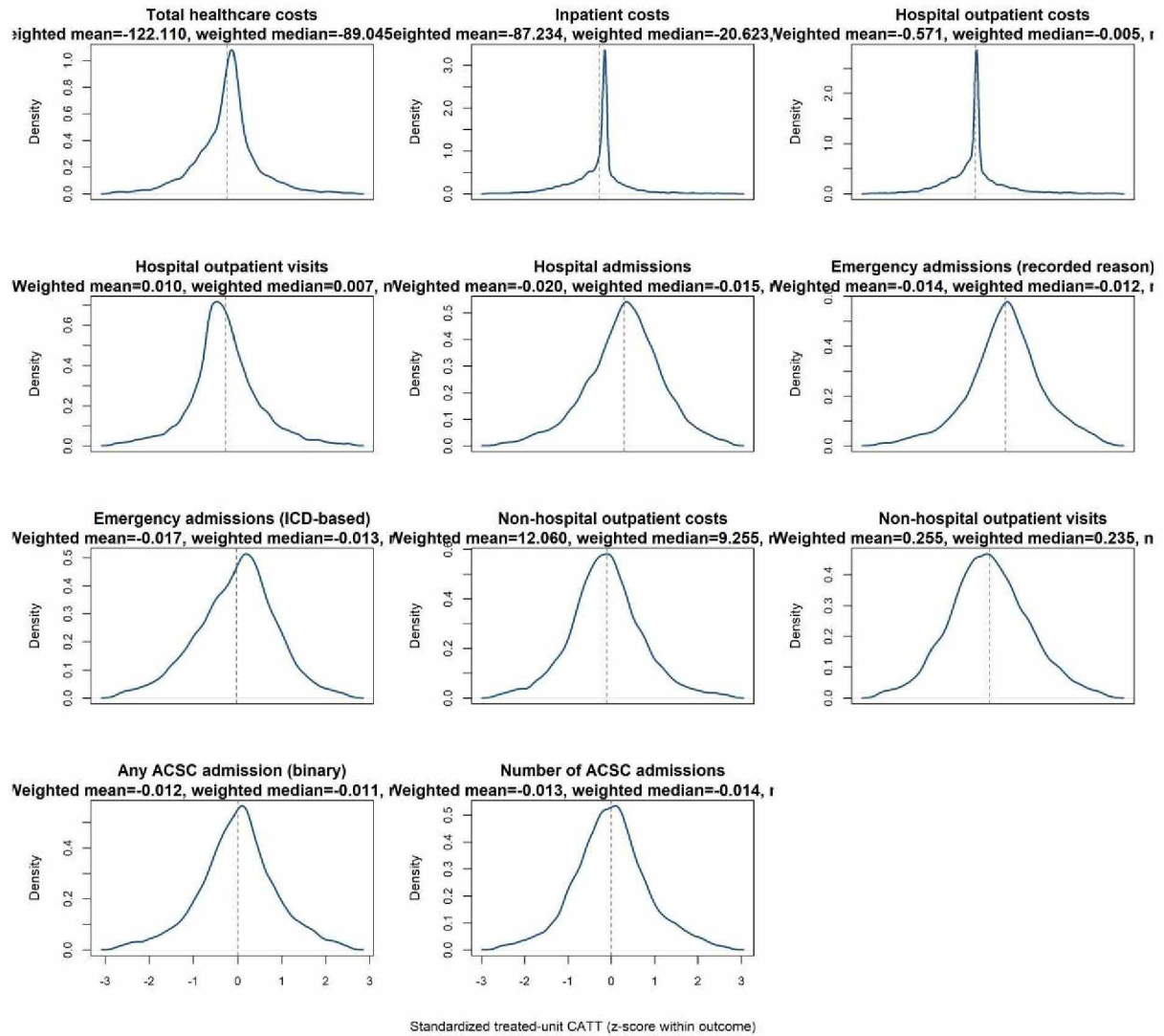
## A.5 Supporting distributional evidence

Figures A4 and A5 provide complementary evidence of effect heterogeneity. The first shows the full standardized distribution of estimated CATTs among treated person-event observations. The second shows how the mean fitted effects of event-specific predicted-effect quartiles evolve over event time.



**Figure A4** Distribution of estimated CATT among treated individuals

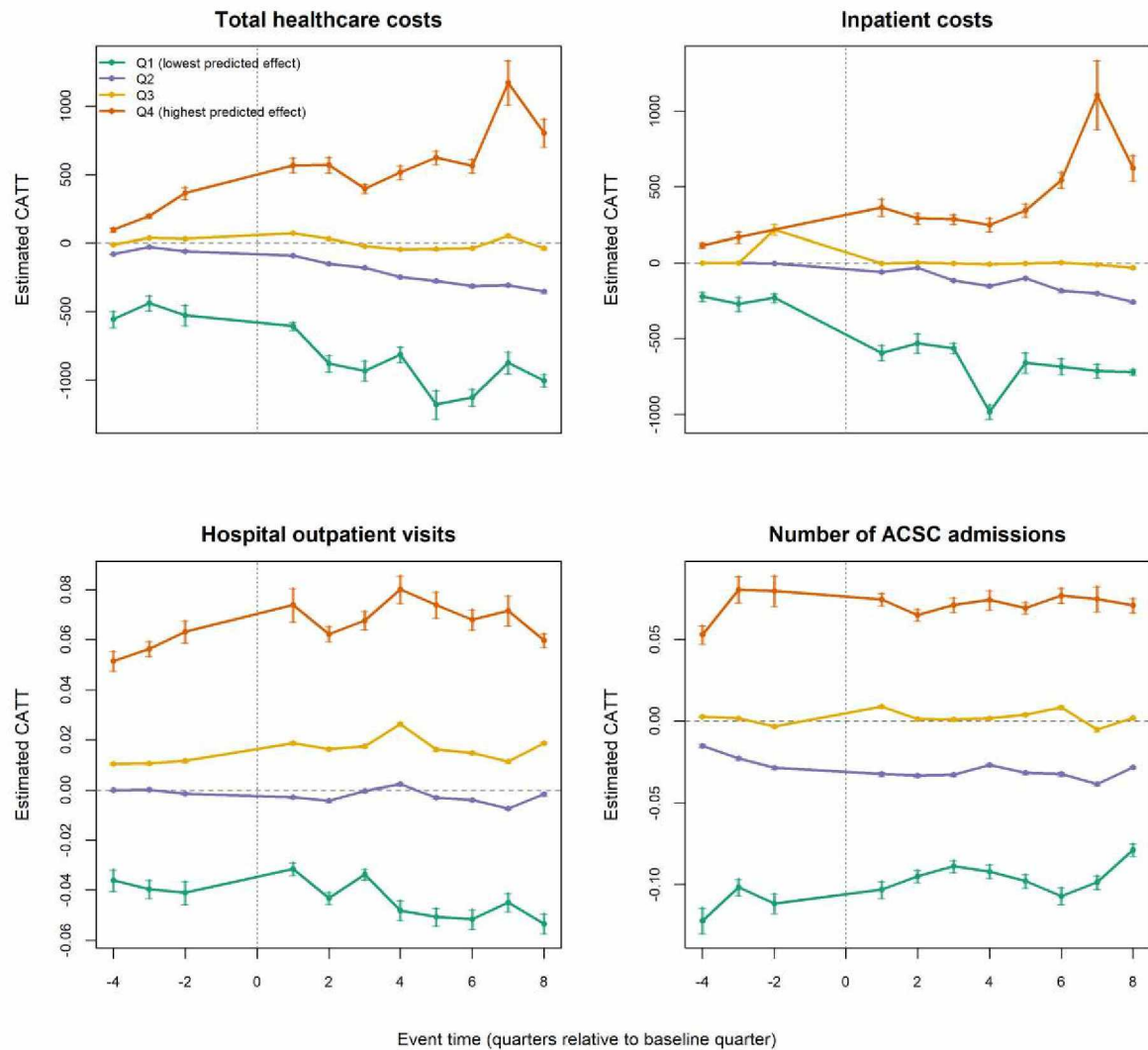
Notes: Curves are sampling-weighted kernel densities of estimated CATTs for treated person-event observations across supported post-enrollment event times  $\ell = 1, \dots, 8$ . Titles report sampling-weighted means and medians;  $n$  is the number of exported treated person-event rows. To limit the influence of extreme tails, densities are displayed between the weighted 1st and 99th percentiles. Panels use outcome-specific scales and original outcome units. The distributions are conditional on the fitted forest and common support; their dispersion combines treatment-effect heterogeneity and estimation uncertainty. Selected outcomes are total healthcare costs, inpatient costs, hospital outpatient visits, and the number of ambulatory care-sensitive condition (ACSC) admissions.



**Figure A5** Distribution of estimated CATTs among treated observations

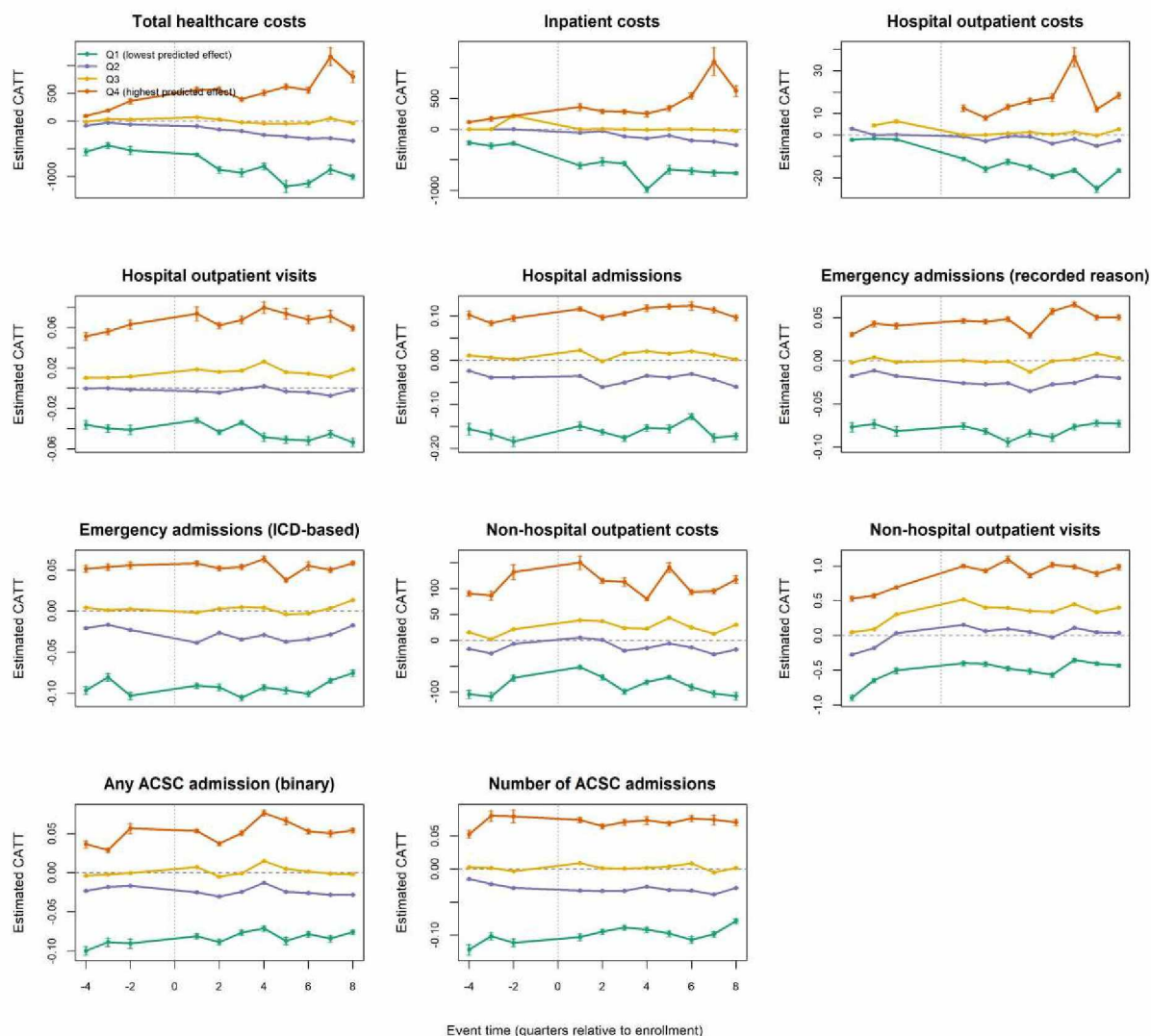
Notes: Curves are sampling-weighted kernel densities of estimated cohort-event-time CATTs for treated person-event observations in supported comparisons. CATTs are standardized within outcome to permit visual comparison; the vertical dashed line marks zero. Densities are displayed between the weighted 1st and 99th percentiles. Panel subtitles report the number of exported treated person-event rows. Dispersion reflects both fitted treatment-effect heterogeneity and estimation uncertainty and should not be interpreted as the distribution of true individual treatment effects.





**Figure A6** Predicted-effect quartiles over event time for selected outcomes

Notes: At each event time separately, treated person–event observations are ranked by predicted CATT and divided into quartiles; Q1 contains the lowest and Q4 the highest predicted effects. Membership can change over time, so lines connect event-specific rank groups rather than fixed patients. Separation is expected by construction and describes its magnitude and evolution rather than independently testing heterogeneity. Points are sampling-weighted quartile means. Bars are pointwise 95% normal-approximation intervals based on within-quartile dispersion and effective sample size, conditional on the fitted forest; they exclude forest-estimation and quartile-assignment uncertainty. Event time is relative to enrollment:  $-1$  is baseline,  $0$  is enrollment and is not included, and positive values are post-enrollment quarters. Panels use outcome-specific scales. ACSC denotes ambulatory care-sensitive conditions.



**Figure A7** Predicted-effect quartiles over event time for all outcomes

Notes: At each event time, treated person-event observations are ranked by their predicted CATT and divided into quartiles; Q1 contains the lowest and Q4 the highest predicted effects. Quartile membership can change over event time, so lines connect event-specific rank groups rather than fixed patients. Separation is expected by construction and summarizes its magnitude and evolution; it is not an independent test of heterogeneity. Points are sampling-weighted quartile means. Bars are pointwise 95% normal-approximation intervals based on within-quartile dispersion and effective sample size, conditional on the fitted forest; they do not include uncertainty from forest estimation or quartile assignment. Panels use outcome-specific scales.



## Appendix B. Additional summaries of heterogeneity

This appendix reports additional summaries of the fitted CATT function. Best linear projections (BLPs) provide linear summaries of how  $\hat{\tau}_{c,\ell}(X_i)$  varies with selected baseline characteristics. Variable-importance measures summarize which groups of covariates contribute most to the forest-estimated heterogeneity. Targeting-oriented diagnostics assess whether the estimated CATT ranking separates patients with larger and smaller predicted responses. These analyses are descriptive summaries of the estimated CATT function and should not be interpreted as causal effects of the baseline covariates themselves.

Across these summaries, heterogeneity is mainly organized by pre-period care history. Prior hospital and ACSC use, outpatient utilization, sector-specific spending, and prior participation in other integrated-care programs repeatedly appear as relevant dimensions of variation. For better-supported inpatient and cost outcomes, larger estimated reductions are observed among patients with greater prior hospital or ACSC use and higher inpatient spending. Patients with little prior acute-care use or no inpatient spending generally show smaller estimated effects.

The BLP results in Tables B2 to B5 provide outcome-specific linear summaries. Prior participation in other integrated-care programs is associated with larger increases in hospital outpatient visits and larger reductions in ACSC admissions, although both estimates are only weakly precise. The total-cost BLP coefficients, including the one for prior integrated-care participation, are imprecisely estimated and do not yield a corresponding pattern. The variable-importance summary in Figure B1 is consistent with this pattern: baseline utilization and cost histories, prior integrated-care contact, medication costs, and inpatient costs are among the most important predictors of heterogeneity across headline outcomes, while sociodemographic covariates are generally secondary.

The targeting-oriented diagnostics suggest that ranking performance is outcome-specific. The ranking signal is more visible for selected inpatient and ACSC outcomes and weaker for total costs and ambulatory outcomes. These diagnostics therefore support heterogeneity in selected endpoints but do not imply a uniform prioritization ranking across all outcomes.

## B.1 Outcome-level summary of estimated heterogeneity

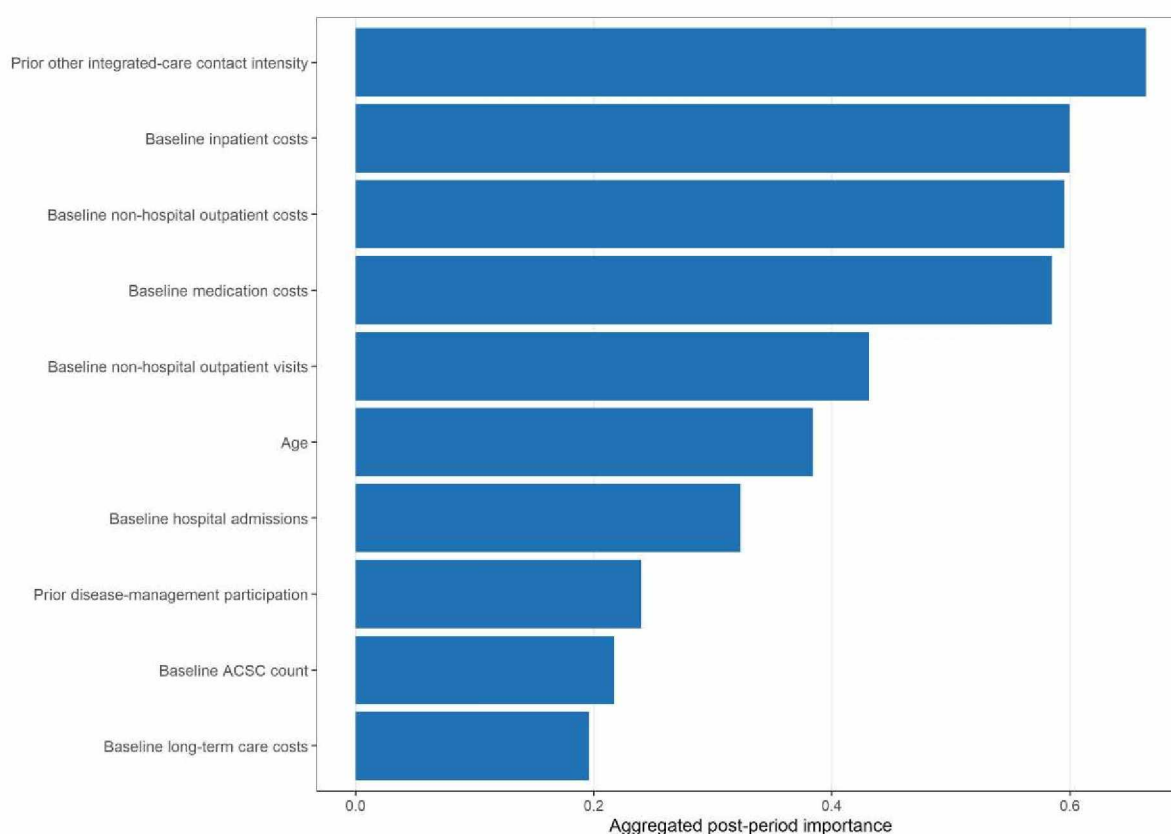
Table B1 summarizes the direction and magnitude of the fitted HTE patterns by outcome. It combines predicted-effect-group contrasts with covariate-binned profiles, BLPs, and variable importance. The diagnostic column is integral to interpretation: a large fitted contrast is not treated as equally credible when pretrend or placebo support is weak.

**Table B1** HTE Summary by Outcome

| Outcome                    | Stronger / more favorable profile                                                      | Weaker / near-zero profile                                                                 | Effect size (95% CI)                                                                     | Main supporting evidence                                                                                                                                                                                                                                                          | Diagnostic caveat                                                                                                             |
|----------------------------|----------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|
| Total healthcare costs     | Higher baseline non-hospital outpatient costs (more negative estimated effects).       | Lower baseline non-hospital outpatient costs (effects closer to zero).                     | At l=1: Q4=543.762 [501.050, 586.474]; Q1=-559.596 [-587.881, -531.312]; Q4-Q1=1103.358. | Top bin-gradient variable: baseline non-hospital outpatient costs (high-low delta=-313.116, Spearman rho=-0.903). BLP for baseline non-hospital outpatient costs: imprecise (p=0.106). Top variable-importance channel: prior care-program participation.                         | pretrend joint Wald p=0.027; placebo joint Wald p=0.085; estimable cell share=0.64; mean  SMD  after adjustment at l=0=0.026. |
| Inpatient costs            | Higher baseline medication costs (more negative estimated effects).                    | Lower baseline medication costs (effects closer to zero).                                  | At l=1: Q4=319.565 [270.653, 368.477]; Q1=-572.259 [-615.669, -528.849]; Q4-Q1=891.824.  | Top bin-gradient variable: baseline medication costs (high-low delta=-268.630, Spearman rho=-0.976). BLP for baseline medication costs: imprecise (p=0.531). Top variable-importance channel: prior care-program participation.                                                   | pretrend joint Wald p=0.789; placebo joint Wald p=0.888; estimable cell share=0.65; mean  SMD  after adjustment at l=0=0.029. |
| Hospital outpatient costs  | Lower age (more negative estimated effects).                                           | Higher age (effects closer to zero or positive).                                           | At l=1: Q4=11.729 [10.445, 13.013]; Q1=-10.266 [-10.968, -9.564]; Q4-Q1=21.995.          | Top bin-gradient variable: age (high-low delta=0.630, Spearman rho=0.867). BLP for age: imprecise (p=0.152). Top variable-importance channel: prior care-program participation.                                                                                                   | pretrend joint Wald p=0.379; placebo joint Wald p=0.895; estimable cell share=0.65; mean  SMD  after adjustment at l=0=0.036. |
| Hospital outpatient visits | Lower prior other integrated-care contact intensity (more negative estimated effects). | Higher prior other integrated-care contact intensity (effects closer to zero or positive). | At l=1: Q4=0.072 [0.066, 0.077]; Q1=-0.032 [-0.034, -0.029]; Q4-Q1=0.103.                | Top bin-gradient variable: prior other integrated-care contact intensity (high-low delta=0.005, Spearman rho=1.000). BLP for prior other integrated-care contact intensity: beta=0.047 [-0.001, 0.096], p=0.057. Top variable-importance channel: baseline utilization and costs. | pretrend joint Wald p=0.295; placebo joint Wald p=0.787; estimable cell share=0.65; mean  SMD  after adjustment at l=0=0.038. |
| Hospital admissions        | Higher baseline non-hospital outpatient costs (more negative estimated effects).       | Lower baseline non-hospital outpatient costs (effects closer to zero).                     | At l=1: Q4=0.114 [0.110, 0.119]; Q1=-0.151 [-0.160, -0.142]; Q4-Q1=0.265.                | Top bin-gradient variable: baseline non-hospital outpatient costs (high-low delta=-0.051, Spearman rho=-0.976). BLP for baseline non-hospital outpatient costs: beta=0.000 [0.000, 0.000], p=0.050. Top variable-importance channel: prior care-program participation.            | pretrend joint Wald p=0.158; placebo joint Wald p=0.187; estimable cell share=0.64; mean  SMD  after adjustment at l=0=0.030. |
| Emergency                  | Higher baseline non-hospital                                                           | Lower baseline non-hospital                                                                | At l=1: Q4=0.045 [0.043, 0.048]; Q1=-                                                    | Top bin-gradient variable: baseline non-hospital                                                                                                                                                                                                                                  | pretrend joint Wald p=0.030; placebo                                                                                          |

|                                   |                                                                                        |                                                                                            |                                                                                      |                                                                                                                                                                                                                                                                                     |                                                                                                                               |
|-----------------------------------|----------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|
| admission s (recorded reason)     | outpatient costs (more negative estimated effects).                                    | outpatient costs (effects closer to zero).                                                 | 0.077 [-0.081, -0.073]; Q4-Q1=0.122.                                                 | outpatient costs (high-low delta=-0.028, Spearman rho=-1.000). BLP for baseline non-hospital outpatient costs: imprecise (p=0.474). Top variable-importance channel: prior care-program participation.                                                                              | joint Wald p=0.017; estimable cell share=0.65; mean  SMD  after adjustment at l=0=0.023.                                      |
| Emergency admission s (ICD-based) | Higher baseline medication costs (more negative estimated effects).                    | Lower baseline medication costs (effects closer to zero).                                  | At l=1: Q4=0.056 [0.053, 0.059]; Q1=-0.089 [-0.092, -0.087]; Q4-Q1=0.145.            | Top bin-gradient variable: baseline medication costs (high-low delta=-0.027, Spearman rho=-0.988). BLP for baseline medication costs: imprecise (p=0.860). Top variable-importance channel: prior care-program participation.                                                       | pretrend joint Wald p=0.047; placebo joint Wald p=0.237; estimable cell share=0.65; mean  SMD  after adjustment at l=0=0.025. |
| Non-hospital outpatient costs     | Lower prior other integrated-care contact intensity (more negative estimated effects). | Higher prior other integrated-care contact intensity (effects closer to zero or positive). | At l=1: Q4=135.851 [125.504, 146.197]; Q1=-53.941 [-57.495, -50.387]; Q4-Q1=189.792. | Top bin-gradient variable: prior other integrated-care contact intensity (high-low delta=18.044, Spearman rho=1.000). BLP for prior other integrated-care contact intensity: imprecise (p=0.450). Top variable-importance channel: prior care-program participation.                | pretrend joint Wald p=0.411; placebo joint Wald p=0.000; estimable cell share=0.65; mean  SMD  after adjustment at l=0=0.030. |
| Non-hospital outpatient visits    | Lower prior other integrated-care contact intensity (more negative estimated effects). | Higher prior other integrated-care contact intensity (effects closer to zero or positive). | At l=1: Q4=0.977 [0.956, 0.998]; Q1=-0.409 [-0.437, -0.380]; Q4-Q1=1.385.            | Top bin-gradient variable: prior other integrated-care contact intensity (high-low delta=0.200, Spearman rho=1.000). BLP for prior other integrated-care contact intensity: beta=0.328 [-0.053, 0.709], p=0.092. Top variable-importance channel: prior care-program participation. | pretrend joint Wald p=0.030; placebo joint Wald p=0.000; estimable cell share=0.63; mean  SMD  after adjustment at l=0=0.025. |
| Any ACSC admission (binary)       | Higher baseline medication costs (more negative estimated effects).                    | Lower baseline medication costs (effects closer to zero).                                  | At l=1: Q4=0.052 [0.050, 0.055]; Q1=-0.082 [-0.085, -0.079]; Q4-Q1=0.134.            | Top bin-gradient variable: baseline medication costs (high-low delta=-0.024, Spearman rho=-0.988). BLP for baseline medication costs: imprecise (p=0.593). Top variable-importance channel: prior care-program participation.                                                       | pretrend joint Wald p=0.000; placebo joint Wald p=0.569; estimable cell share=0.66; mean  SMD  after adjustment at l=0=0.027. |
| Number of ACSC admission s        | Higher baseline medication costs (more negative estimated effects).                    | Lower baseline medication costs (effects closer to zero).                                  | At l=1: Q4=0.071 [0.068, 0.075]; Q1=-0.106 [-0.110, -0.101]; Q4-Q1=0.177.            | Top bin-gradient variable: baseline medication costs (high-low delta=-0.034, Spearman rho=-0.988). BLP for baseline medication costs: imprecise (p=0.750). Top variable-importance channel: prior care-program participation.                                                       | pretrend joint Wald p=0.137; placebo joint Wald p=0.301; estimable cell share=0.65; mean  SMD  after adjustment at l=0=0.024. |

Notes: The table summarizes heterogeneity in estimated cohort-event-time conditional average treatment effects on the treated. “Stronger/more favorable” and “weaker/near-zero” profiles are based first on predicted-effect quartiles and subgroup contrasts, and then on best linear projections, variable-importance rankings, targeting diagnostics, and robustness checks where available. Effect sizes are reported in the original outcome units: EUR for costs, counts for utilization outcomes, and percentage points for binary outcomes. Variable importance is descriptive of the fitted forest and is not interpreted as causal evidence that the covariate itself changes treatment effects. Subgroup claims are interpreted only within supported cohort-event-time comparisons and should be read together with the diagnostic caveats.



**Figure B1** Most important baseline covariates

Note: Aggregated variable importance across all outcomes.

**Table B2** Best linear projection of estimated CATTs: number of ACSC admissions

| Baseline characteristic        | Coefficient       |
|--------------------------------|-------------------|
| ACSC count                     | -0.154<br>(0.180) |
| ACSC history, any              | 0.246<br>(0.767)  |
| Hospital admissions            | 0.018<br>(0.072)  |
| Non-hospital ambulatory visits | 0.001<br>(0.006)  |
| Integrated-care contact        | -0.060<br>(0.036) |
| Observations used              | 44,986            |

Notes: BLP = best linear projection. The dependent variable is the fitted cohort-eventtime CATT among treated observations. Standard errors are shown in parentheses. Coefficients are in original outcome units. Covariates are selected to be outcome-

relevant for each separate portrait table; the combined appendix BLP table uses the same prespecified covariates across outcomes for comparability. BLP coefficients summarize associations between baseline characteristics and the fitted treatment-effect function; they are not causal effects of the covariates themselves. Significance stars: \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.10$ .

**Table B3** Best linear projection of estimated CATTs: inpatient costs

| Baseline characteristic | Coefficient             |
|-------------------------|-------------------------|
| Inpatient costs         | -0.829<br>(0.597)       |
| Hospital admissions     | -7.250<br>(359.576)     |
| ACSC history, any       | -1725.847<br>(2776.518) |
| Integrated-care contact | 305.462<br>(366.024)    |
| Observations used       | 45,000                  |

Notes: BLP = best linear projection. The dependent variable is the fitted cohort-eventtime CATT among treated observations. Standard errors are shown in parentheses. Coefficients are in original outcome units. Covariates are selected to be outcome-relevant for each separate portrait table; the combined appendix BLP table uses the same prespecified covariates across outcomes for comparability. BLP coefficients summarize associations between baseline characteristics and the fitted treatment-effect function; they are not causal effects of the covariates themselves. Significance stars: \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.10$ .

**Table B4** Best linear projection of estimated CATTs: total healthcare costs

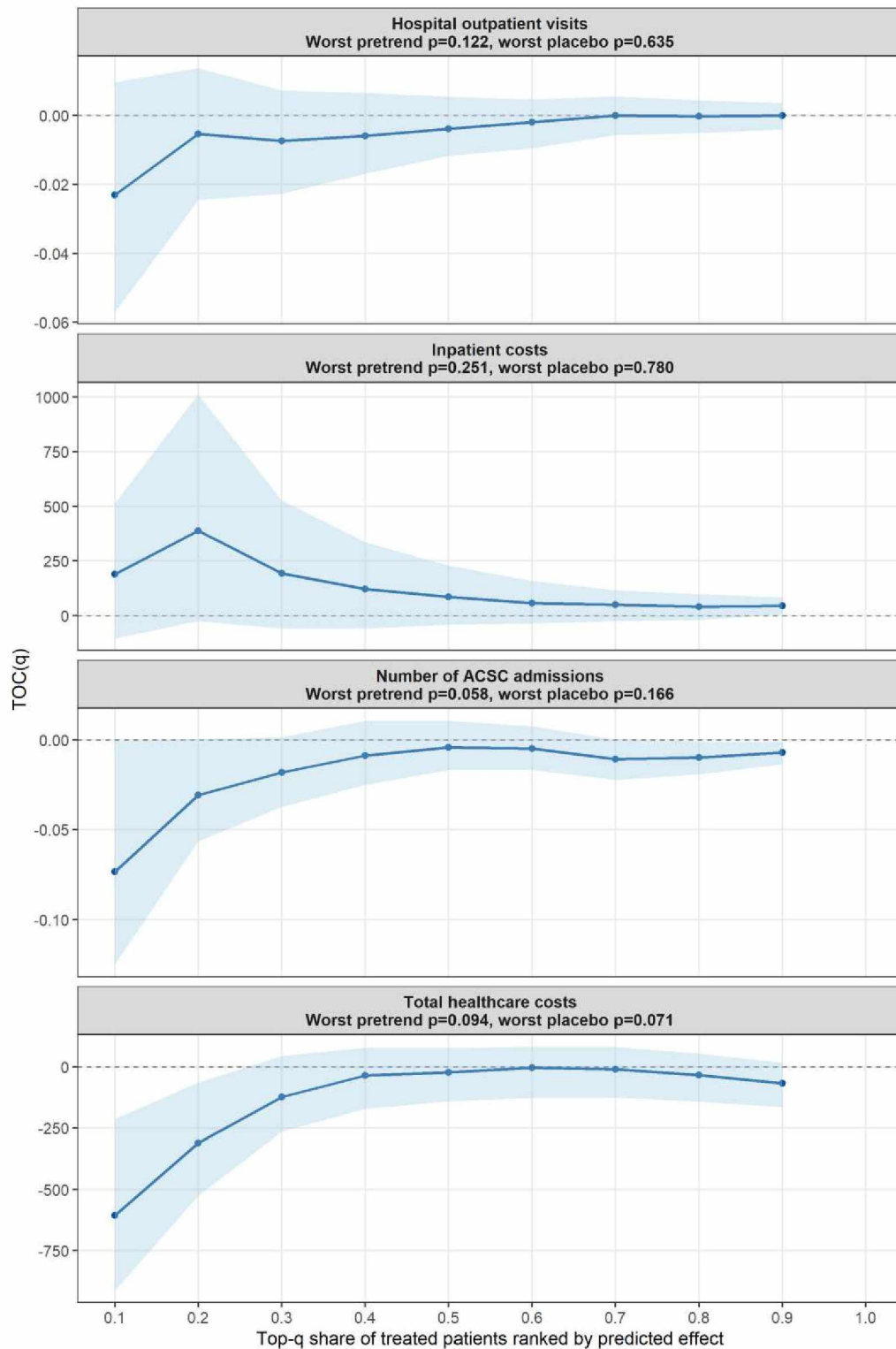
| Baseline characteristic | Coefficient           |
|-------------------------|-----------------------|
| Inpatient costs         | -0.637<br>(0.907)     |
| Hospital admissions     | -635.370<br>(566.866) |
| ACSC history, any       | 636.120<br>(4326.894) |
| Integrated-care contact | 101.282<br>(452.446)  |
| Observations used       | 45,300                |

Notes: BLP = best linear projection. The dependent variable is the fitted cohort-eventtime CATT among treated observations. Standard errors are shown in parentheses. Coefficients are in original outcome units. Covariates are selected to be outcome-relevant for each separate portrait table; the combined appendix BLP table uses the same prespecified covariates across outcomes for comparability. BLP coefficients summarize associations between baseline characteristics and the fitted treatment-effect function; they are not causal effects of the covariates themselves. Significance stars: \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.10$ .

**Table B5** Best linear projection of estimated CATTs: hospital outpatient visits

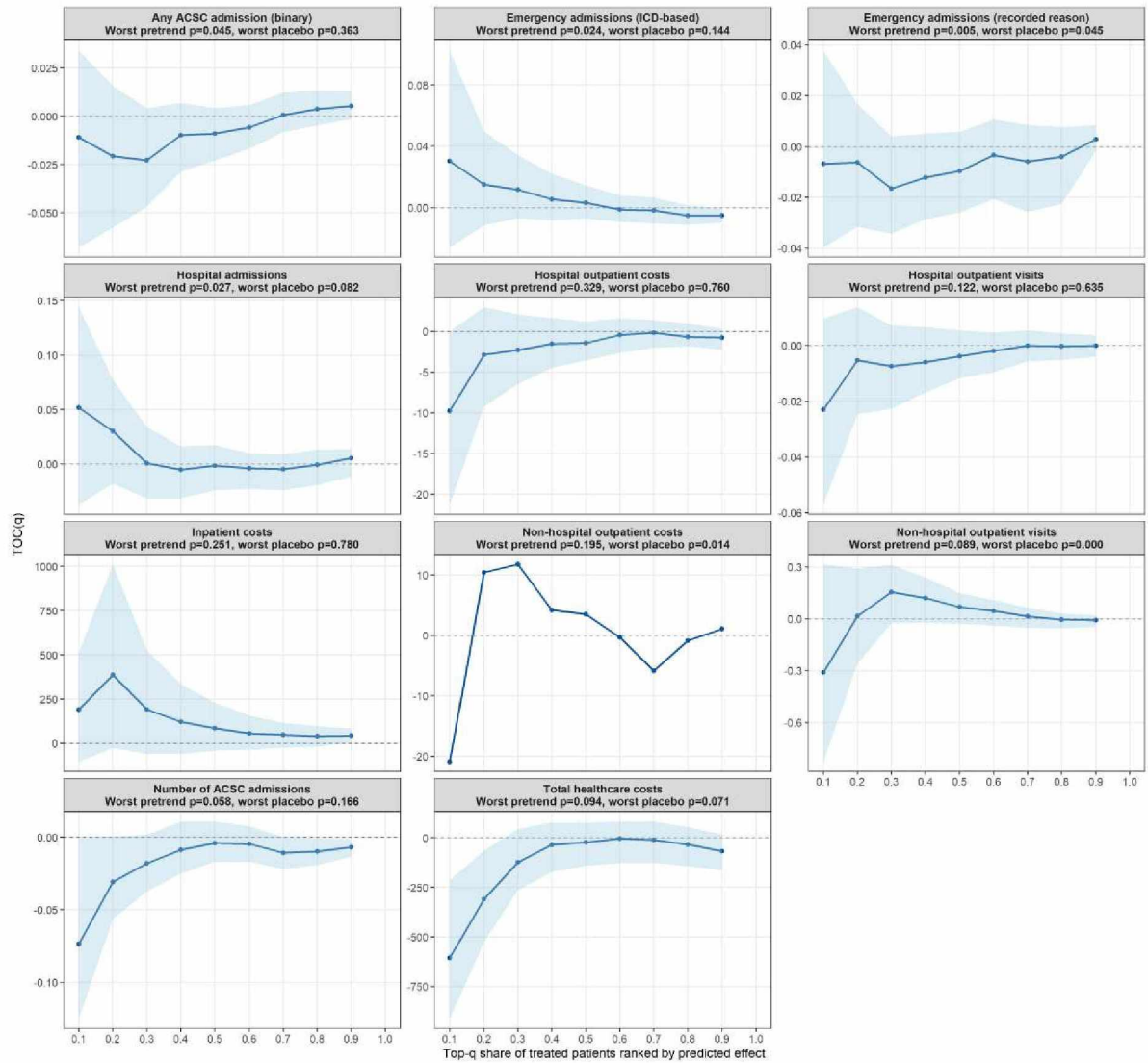
| Baseline characteristic          | Coefficient       |
|----------------------------------|-------------------|
| Hospital outpatient costs        | 0.000<br>(0.000)  |
| Non-hospital ambulatory visits   | -0.004<br>(0.004) |
| Hospital admissions              | 0.017<br>(0.102)  |
| Disease-management participation | -0.005<br>(0.021) |
| Integrated-care contact          | 0.045*<br>(0.024) |
| Observations used                | 44,700            |

Notes: BLP = best linear projection. The dependent variable is the fitted cohort-eventtime CATT among treated observations. Standard errors are shown in parentheses. Coefficients are in original outcome units. Covariates are selected to be outcome-relevant for each separate portrait table; the combined appendix BLP table uses the same prespecified covariates across outcomes for comparability. BLP coefficients summarize associations between baseline characteristics and the fitted treatment-effect function; they are not causal effects of the covariates themselves. Significance stars: \*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.10$ .



**Figure B2** Targeting Operator Characteristic (TOC) Curves for Predicted CATT Rankings

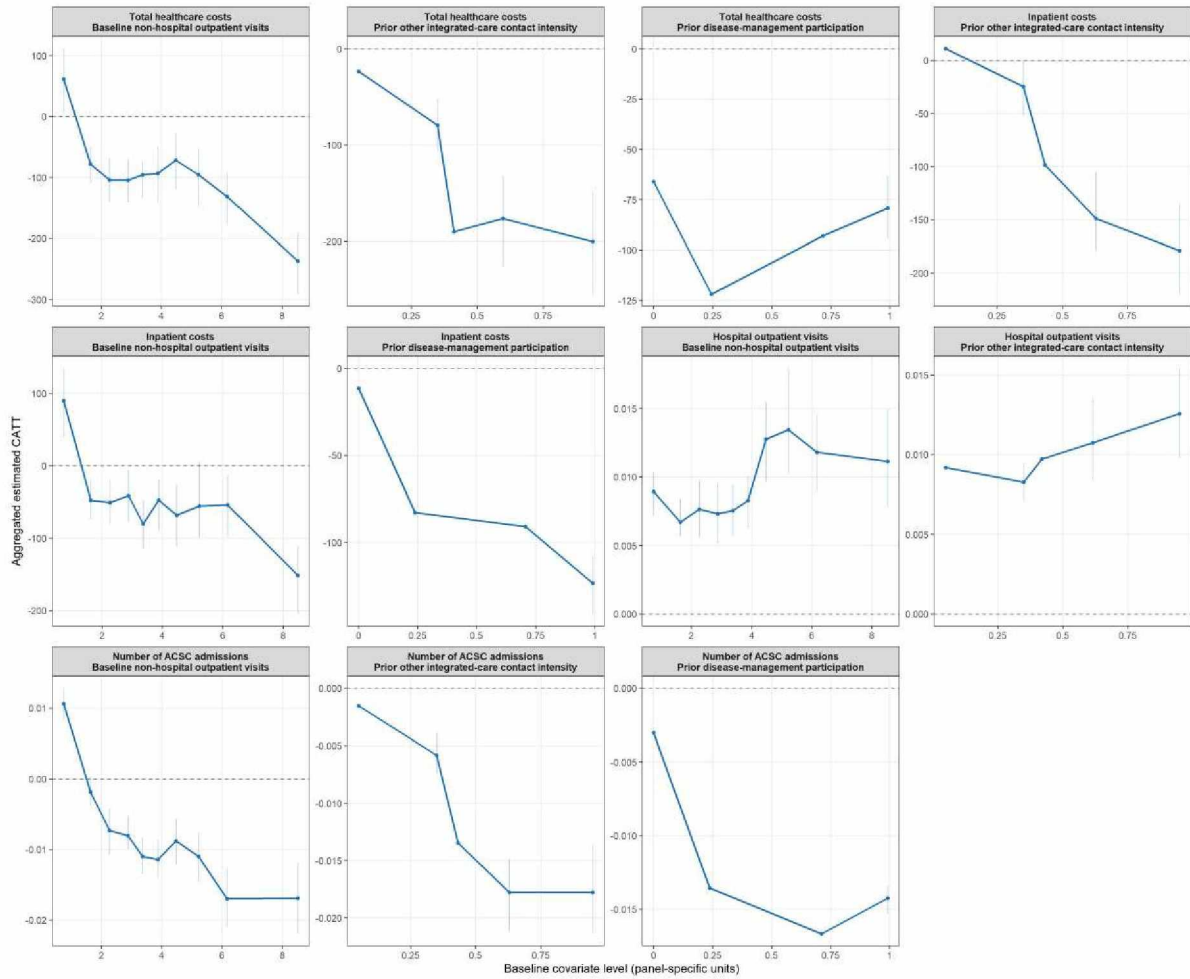
Notes: TOC( $q$ ) is the sampling-weighted mean doubly robust score among the top- $q$  share of treated person–event observations ranked from highest to lowest predicted CATT, minus the corresponding mean for all treated observations. Scores are pooled over supported post-enrollment event times 1–8. Positive values indicate above-average effects in the selected share; negative values indicate below-average effects. Because ranking is by the numerical CATT rather than benefit, interpretation is reversed when reductions are preferred, as for costs and admissions. Shaded areas are pointwise 95% bootstrap intervals from 120 unit-score resamples where available, conditional on the fitted forest. Panels use outcome-specific scales. TOC denotes targeting operator characteristic; ACSC denotes ambulatory care-sensitive conditions.



**Figure B3** Targeting Operator Characteristic (TOC) Curves for Predicted CATT Rankings (all Outcomes)

Notes: TOC( $q$ ) is the sampling-weighted mean doubly robust score among the top- $q$  share of treated person–event observations ranked from highest to lowest predicted CATT, minus the corresponding mean for all treated observations. Scores are pooled over supported post-enrollment event times  $\ell = 1, \dots, 8$ . Positive values indicate above-average effects in the selected share; negative values indicate below-average effects. Because ranking is by the numerical CATT rather than benefit, interpretation is reversed when reductions are preferred, as for costs and admissions. Shaded areas are pointwise 95% bootstrap intervals from 120 unit-score resamples where available, conditional on the fitted forest. Panels use outcome-specific scales. TOC denotes targeting operator characteristic; ACSC denotes ambulatory care-sensitive conditions.





**Figure B4** Estimated HTEs by baseline covariate level for all outcomes

Notes: Points are weighted mean predicted CATTs among treated observations, pooled across post-enrollment event times  $\ell = 1, \dots, 8$ , within empirical covariate bins. Horizontal positions show mean covariate values within bins. Unless otherwise stated, baseline utilization and spending measures are averages per quarter over the four quarters preceding enrollment. Prior other integrated-care contact intensity is the share of all observed pre-enrollment quarters with participation in another integrated-care contract. For each outcome, the figure shows the two highly ranked covariates with the largest range in binned mean CATTs. Bars are pointwise 95% individual-cluster bootstrap intervals where available, conditional on the fitted forest. Baseline outpatient visits are four-quarter pre-enrollment means. ACSC denotes ambulatory care-sensitive conditions.

## Appendix C. Robustness and sensitivity analyses

This appendix reports robustness checks for the main staggered DiD causal-forest specification. The checks alter the analysis sample, treatment timing, common-support rules, outcome construction, imputation, and clustering. Because several checks change the estimand or the available cohort-event-time comparisons, the evidence is evaluated using changes in point estimates, uncertainty, diagnostic support, and estimable coverage.

### C.1 Coverage and diagnostic support

**Table C1** Robustness and sensitivity checks

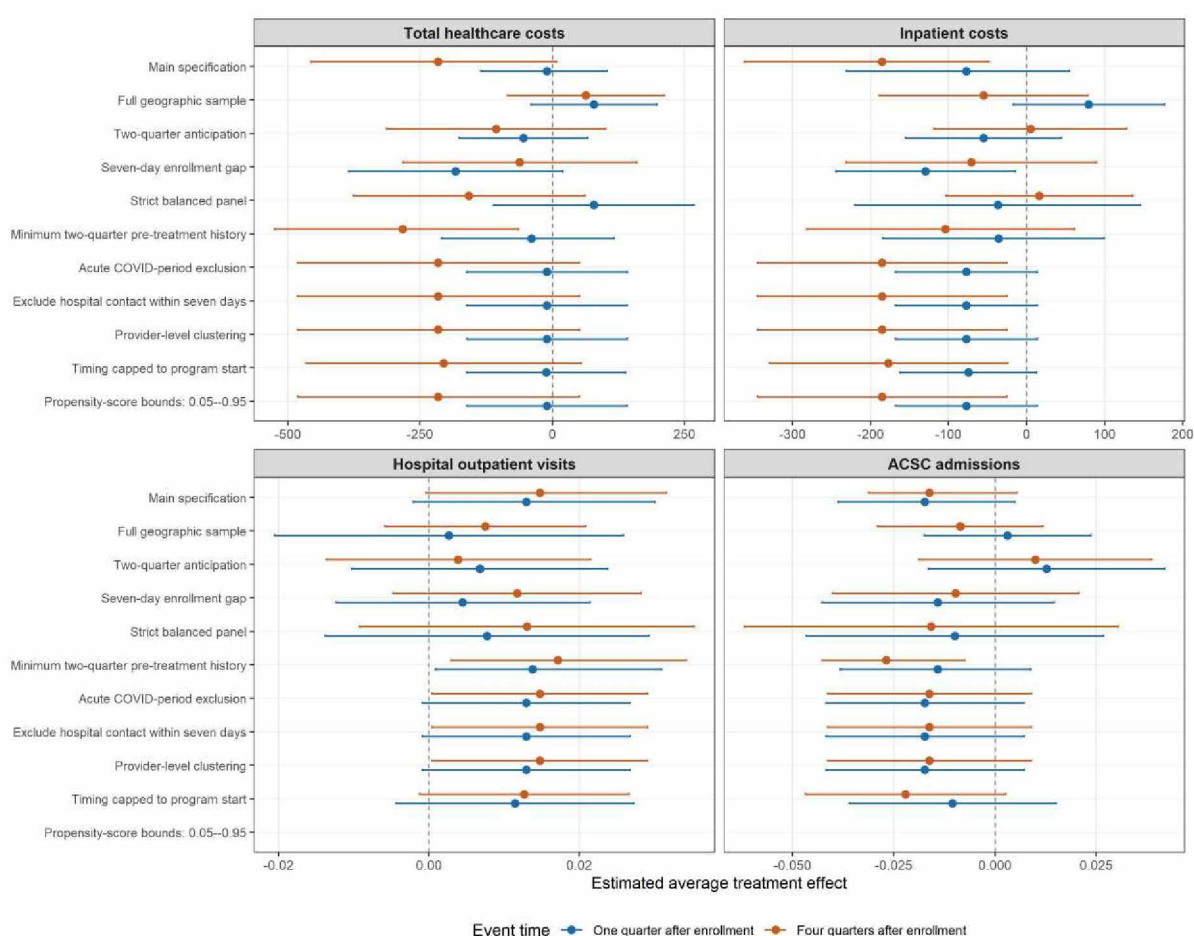
| Specification                             | Dimension changed                                                                     | Median cell share | Interpretation                                                             |
|-------------------------------------------|---------------------------------------------------------------------------------------|-------------------|----------------------------------------------------------------------------|
| Main specification                        | Target-area adults; 0.03–0.97 propensity-score bounds                                 | 64.9%             | Supports comparison with the main specification for the selected outcomes. |
| Full geographic sample                    | Includes insured persons outside the target-area analysis sample                      | 80.4%             | Partly informative, but diagnostic support varies by outcome.              |
| Two-quarter anticipation                  | Moves the reference period back by two additional quarters                            | 59.4%             | Estimable, but diagnostic <i>p</i> -values are unavailable.                |
| Seven-day enrollment gap                  | Alternative treatment-timing rule around enrollment                                   | 60.6%             | Partly informative, but diagnostic support varies by outcome.              |
| Strict balanced panel                     | Restricts estimation to a balanced observation window                                 | 51.1%             | Supports comparison with the main specification for the selected outcomes. |
| Minimum two-quarter pre-treatment history | Requires at least two observed quarters within the four-quarter pre-treatment history | 64.7%             | Supports comparison with the main specification for the selected outcomes. |
| Acute COVID-period exclusion              | Excludes prespecified acute-pandemic calendar quarters                                | 64.9%             | Supports comparison with the main specification for the selected outcomes. |

|                                              |                                                                                        |       |                                                                            |
|----------------------------------------------|----------------------------------------------------------------------------------------|-------|----------------------------------------------------------------------------|
| Exclude same-day hospital contact            | Drops treated persons linked to hospital care on enrollment date                       | 64.9% | Supports comparison with the main specification for the selected outcomes. |
| Exclude enrollment-quarter hospital contact  | Drops treated persons linked to hospital care during enrollment                        | 64.9% | Supports comparison with the main specification for the selected outcomes. |
| Exclude hospital contact within seven days   | Drops treated persons linked to hospital care within seven days                        | 64.9% | Supports comparison with the main specification for the selected outcomes. |
| Provider-level clustering                    | Clusters inference by pre-treatment provider                                           | 64.9% | Supports comparison with the main specification for the selected outcomes. |
| Timing capped to program start               | Caps recorded timing at the program start quarter                                      | 64.9% | Partly informative, but diagnostic support varies by outcome.              |
| No trimming or cell-level overlap exclusions | Removes 0.03–0.97 trimming and disables the common-support-share and KS-distance rules | 64.9% | Supports comparison with the main specification for the selected outcomes. |
| Propensity-score bounds: 0.05–0.95           | Tightens the main 0.03–0.97 bounds while retaining the cell-level overlap rules        | 64.9% | Supports comparison with the main specification for the selected outcomes. |
| Real-term costs                              | Adjusts cost outcomes for calendar-time price changes                                  | 64.9% | Supports comparison with the main specification for the selected outcomes. |
| Cost winsorization                           | Winsorizes cost outcomes at the 1st and 99th percentiles                               | 64.9% | Partly informative, but diagnostic support varies by outcome.              |

Notes: Selected outcomes are total healthcare costs, inpatient costs, hospital outpatient visits, and ACSC admissions. Diagnostics passed reports the number of selected outcomes with synchronized estimable results for which both the worst-pretrend and placebo p-values are at least 0.05. Cell share is the number of estimated cohort-event-time comparisons divided by all candidate comparisons. The main specification trims propensity scores outside 0.03–0.97 and applies the common-support-share and KS-distance rules. The no-exclusion check removes these three restrictions but retains the minimum treated and comparison sample-size requirements. Missing synchronized selected-outcome results are not evidence of no effect; some checks were also run for additional outcomes not shown in this selected-outcome table.

## C.2 Event-time estimates

Figure C1 and Table C2 compare estimates one and four quarters after enrollment. The evidence is not uniform across all variants. The full geographic sample changes the early cost estimates and has weaker placebo support for total costs and hospital outpatient visits. The two-quarter anticipation and strict balanced-panel specifications also alter some inpatient and ACSC estimates. Requiring at least two observed pre-treatment quarters preserves the signs of the selected estimates but changes their magnitude and precision, especially at four quarters. The seven-day timing rule preserves several signs but fails the pretrend threshold for total costs, inpatient costs, and ACSC admissions. By contrast, the acute-COVID exclusion, hospital-contact exclusion, provider-clustering, and timing-cap checks leave the main point-estimate pattern largely unchanged in the synchronized outputs. These results support the headline interpretation selectively rather than demonstrating invariance across all specifications.

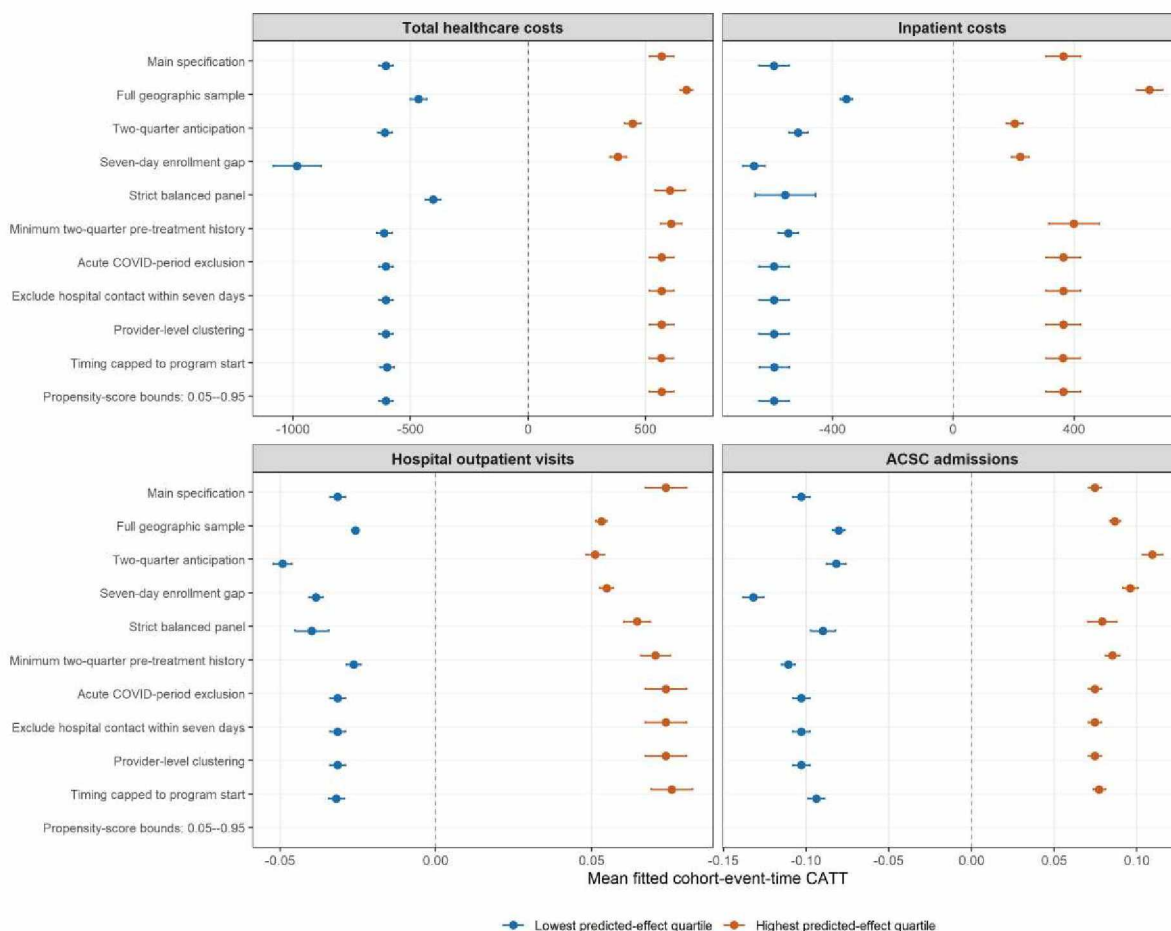


**Figure C1** Selected event-time estimates across robustness specifications

Notes: Points report average treatment effects at 1 and 4 quarters after enrollment; horizontal bars indicate 95% confidence intervals. Panels use outcome-specific scales. Specifications that change timing, samples, or support may target different treated populations.

### C.3 Heterogeneity across specifications

Figure C2 compares the lowest and highest predicted-effect quartiles one quarter after enrollment. Separation remains visible across most supported variants, especially for the cost outcomes, but its magnitude and level vary when the geographic sample, treatment timing, or observation window changes. The quartiles are re-estimated within each specification. The figure therefore shows whether fitted CATT dispersion persists; it does not show that the same patients remain in the same quartile.



**Figure C2** Predicted-effect quartiles across robustness specifications

Notes: Points are mean fitted cohort-event-time CATTs in the lowest and highest predicted-effect quartiles one quarter after enrollment; horizontal bars are the run-provided 95% intervals. Patients are ranked again within each specification, so quartile membership may change. The figure assesses persistence of fitted heterogeneity, not stability of individual rankings. Panels use outcome-specific scales.

## Appendix D: The Difference-in-Differences Causal Forest Method

This appendix provides additional technical detail on the Difference-in-Differences Causal Forest (DiD-CF) estimator used in the main analysis (Gavrilova et al., 2025). It complements the empirical framework in Section 5. The goal is to make explicit how staggered treatment timing, baseline differencing, nuisance adjustment, forest estimation, and support restrictions enter the estimation of cohort-event-time conditional average treatment effects on the treated (CATTs).

### D.1 Set-up and estimand

Individuals enter the integrated-care advice and navigation service in different calendar quarters. Let  $G_i$  denote the first enrollment quarter of individual  $i$ , with  $G_i = \infty$  for individuals who are never enrolled during the observation window. Let  $c$  index enrollment cohorts, where cohort  $c$  consists of individuals first enrolled in calendar quarter  $g_c$ . Event time is defined relative to enrollment,

$$\ell = t - g_c,$$

so that  $\ell = 0$  is the enrollment quarter. The estimator allows for an anticipation buffer of  $a \geq 0$  quarters, so that the cohort-specific baseline quarter is  $b_c = g_c - a - 1$  and outcomes in the  $a$  quarters immediately preceding enrollment are not used as the untreated reference period. The reported results use  $a = 0$ , so  $b_c = g_c - 1$ , corresponding to  $\ell = -1$ ;  $a$  is a configurable design choice, not a fixed feature of the estimator.

For outcome  $Y_{it}$ , the baseline-differenced outcome in cohort-event-time comparison  $(c, \ell)$  is

$$z_{i,c,\ell} = Y_{i,g_c+\ell} - Y_{i,b_c}.$$

For the baseline quarter itself,  $z_{i,c,-1} = 0$  by construction. The treatment indicator in comparison  $(c, \ell)$  is

$$W_i^{c,\ell} = 1\{G_i = g_c\},$$

which equals one for individuals in enrollment cohort  $c$  and zero for valid untreated comparison observations.

The valid untreated comparison set is

$$U_{c,\ell} = \{i: G_i = \infty \text{ or } G_i > g_c + \ell + a\}.$$

Thus, comparison observations are either never treated or remain untreated through  $a$  quarters beyond the calendar quarter corresponding to event time  $\ell$ ; with the reported  $a = 0$ , this is simply not yet treated in that quarter. Already-treated individuals are not used as controls. This avoids comparisons between newly treated and already treated individuals, which can contaminate staggered treatment designs when effects vary across cohorts or event time.

The primary estimand is the cohort-event-time CATT,

$$\tau_{c,\ell}(x) = E[Y_{i,g_c+\ell}(1) - Y_{i,g_c+\ell}(0) \mid G_i = g_c, X_i = x],$$

where  $X_i$  denotes predetermined baseline covariates. Because the design identifies effects from changes relative to the cohort-specific baseline quarter, the estimand is equivalently written in terms of  $z_{i,c,\ell}$  as

$$\tau_{c,\ell}(x) = E[z_{i,c,\ell}(1) \mid W_i^{c,\ell} = 1, X_i = x] - E[z_{i,c,\ell}(0) \mid W_i^{c,\ell} = 1, X_i = x].$$

The second term is counterfactual for treated individuals and must be identified using valid untreated comparison observations.

## D.2 Identification assumptions

The key identifying assumption is conditional parallel trends in untreated outcome changes. For each estimable cohort-event-time comparison  $(c, \ell)$  and covariate value  $x$ ,

$$E[z_{i,c,\ell}(0) \mid W_i^{c,\ell} = 1, X_i = x] = E[z_{i,c,\ell}(0) \mid W_i^{c,\ell} = 0, X_i = x, i \in U_{c,\ell}].$$

This assumption allows enrolled and comparison individuals to differ in baseline morbidity, utilization, and costs. It requires that, after conditioning on observed pre-treatment histories and within common support, their untreated outcome changes would have been comparable.

The design also requires no anticipation, overlap, and no interference. No anticipation means that outcomes are not affected before enrollment by future participation; the estimator operationalizes this via the anticipation buffer  $a$  introduced above;  $a = 0$  in the reported results, so the assumption is invoked only for the single quarter immediately preceding enrollment ( $b_c$  itself). A nonzero  $a$  would push the reference period further back and additionally require comparison units to remain untreated for  $a$  quarters beyond event time  $\ell$ , at the cost of a shorter usable pre-period. Overlap requires that treated individuals in cohort  $c$  have comparable untreated observations with similar covariates in the relevant comparison  $(c, \ell)$ . No interference means that one individual's potential outcomes are not directly affected by another individual's enrollment status.

Under conditional parallel trends and overlap, the counterfactual untreated change for treated individuals can be identified from valid untreated comparison observations. Hence,

$$\tau_{c,\ell}(x) = E[z_{i,c,\ell} \mid W_i^{c,\ell} = 1, X_i = x] - E[z_{i,c,\ell} \mid W_i^{c,\ell} = 0, X_i = x, i \in U_{c,\ell}].$$

This is the DiD analogue of a conditional treatment effect: the forest learns heterogeneity in outcome changes, not heterogeneity in outcome levels.

## D.3 Orthogonalized forest estimation within a cohort-event-time comparison

For each cohort-event-time comparison  $(c, \ell)$ , estimation is performed on the sample consisting of treated individuals in cohort  $c$  and valid untreated comparison observations in  $U_{c,\ell}$ . To simplify notation in this subsection, write  $z_i = z_{i,c,\ell}$  and  $W_i = W_i^{c,\ell}$ .

The DiD-CF estimator uses nuisance functions for the baseline-differenced outcome and treatment-cohort membership:

$$m_{c,\ell}(X_i) = E[z_i \mid X_i], \quad p_{c,\ell}(X_i) = P(W_i = 1 \mid X_i).$$



The causal forest is estimated using residualized outcomes and residualized treatment indicators. The orthogonal moment condition can be written as

$$E[\{W_i - p_{c,\ell}(X_i)\}\{Z_i - m_{c,\ell}(X_i) - \tau_{c,\ell}(X_i)(W_i - p_{c,\ell}(X_i))\} | X_i = x] = 0.$$

This moment makes the treatment-effect estimate less sensitive to first-stage errors in the outcome and propensity models. The nuisance functions are not the substantive object of interest; they are used to adjust for observable differences in baseline histories before estimating how treatment effects vary with  $X_i$ .

In implementation, estimation within a cohort-event-time comparison proceeds in two stages, reflecting that propensity trimming (Section 1.6) is itself based on a fitted propensity model.

In the first stage, candidate nuisance models are compared by two-fold cross-validation: the propensity model is chosen between a logistic and a LASSO-penalized logistic specification, and the outcome model between a linear and an elastic-net-penalized specification, selecting in each case the candidate with the better two-fold out-of-fold performance (log-loss for propensity, mean squared error for the outcome model). This selection step uses genuine cross-fitting: each observation's nuisance prediction used for model comparison is generated by a model trained on the other fold. A causal forest is then estimated on the full (untrimmed) comparison sample using these cross-fitted nuisance predictions, and its propensity estimates are the ones used for the trimming rule below.

In the second stage, after trimming, the previously selected propensity and outcome model *types* are refit on the trimmed sample only, and predictions are generated in-sample (i.e., on the same trimmed observations used to fit them) rather than by a further round of cross-fitting. A causal forest is then re-estimated on the trimmed sample using these in-sample nuisance predictions, and this second forest is the one whose treatment-effect estimates, heterogeneity predictions, and diagnostics are reported throughout the paper. This means the out-of-sample property that motivates cross-fitting – and that the orthogonal moment condition above is designed to exploit – holds for the nuisance predictions used to decide trimming, but not for the nuisance predictions used in the forest that produces the reported estimates. This is a design trade-off (avoiding a second, more expensive cross-fitting pass per cell) rather than an oversight, but it means the reported CATTs rely on the causal forest's own robustness to first-stage nuisance error and honest splitting for bias control, rather than on out-of-sample nuisance prediction alone.

#### D.4 Forest weights and relation to matching

The causal forest partitions the covariate space into leaves and estimates local treatment effects using observations that are similar in baseline covariates. For a treated individual with covariates  $x$ , the forest can be viewed as assigning weights to observations whose covariates frequently fall in the same terminal leaves as  $x$  across trees. Observations with more similar baseline profiles receive greater weight.

This creates a matching-like intuition, but the estimator is not a conventional matching estimator. We do not select one-to-one matches, impose calipers, or construct a matched sample. Propensity scores enter the orthogonal score and support diagnostics; they do not define matched pairs. Comparability is addressed through the conditional DiD design, common-support restrictions, residualization, forest-based local weighting, and balance diagnostics.



The forest is estimated honestly. Within each tree, one subsample determines the partition of the covariate space, while a separate subsample estimates effects within the resulting leaves. Honesty reduces overfitting because the same observations are not used both to choose the splits and to estimate the treatment effect in the leaves.

### D.5 Staggered adoption and aggregation

The procedure is repeated across cohort-event-time comparisons. Each comparison has its own treated cohort, event-time quarter, valid untreated comparison set, nuisance functions, support diagnostics, and forest estimate. The estimator therefore produces CATT predictions

$$\hat{\tau}_{c,\ell}(X_i)$$

for treated individuals in estimable comparisons.

The scalar ATT reported for each cohort-event-time comparison is not the simple average of the unit-level CATT predictions  $\hat{\tau}_{c,\ell}(X_i)$ ; it is the forest’s doubly-robust (augmented inverse-propensity-weighted) treated-sample average-treatment-effect estimator, which combines the forest’s CATT predictions with the fitted propensity and outcome nuisances so that the reported cell-level ATT retains some protection against forest misspecification even if the CATT surface itself is imperfectly estimated.

Average dynamic effects at a given event time are obtained by aggregating these cohort-event-time ATT estimates across cohorts, weighted by each cohort’s number of treated observations in that cell: cohorts contributing more treated individuals to a given event time receive proportionally more weight, rather than every contributing cohort being weighted equally. These averages are useful benchmarks, but they are not the central object of the paper. The main object is the distribution and structure of  $\hat{\tau}_{c,\ell}(X_i)$  among enrolled individuals.

Heterogeneity is summarized in several complementary ways. First, predicted-effect distributions show the dispersion of estimated CATTs among treated observations. Second, predicted-effect quartiles sort treated observations by their estimated  $\hat{\tau}_{c,\ell}(X_i)$  and compare average estimated effects across the distribution of predicted effects. Third, covariate-binned profiles summarize how estimated CATTs vary across baseline patient characteristics such as prior utilization, costs, ACSC history, and prior other integrated-care contact. Fourth, best linear projections provide low-dimensional summaries of how fitted CATTs co-vary with selected baseline covariates. Variable-importance measures describe which covariates the forest used most often for heterogeneity, but they are interpreted descriptively and not as causal effects of the covariates themselves.

### D.6 Support restrictions and estimability

The DiD-CF estimator is used only where the data support treated-comparison contrasts. In the main specification, observations are trimmed when the treatment propensity estimated by the first-stage (untrimmed-sample) forest is below 0.03 or above 0.97. After trimming and re-estimation, a cohort-event-time comparison is retained only if, according to the second-stage (post-trim) forest’s propensity estimates, at least 60 percent of observations remain within common support and the Kolmogorov-Smirnov distance between the treated and comparison propensity-score distributions does not exceed 0.80. We also require at least 40 treated observations and 120 comparison observations, checked both immediately after trimming and again after these overlap checks.

These restrictions are design choices rather than technical failures. They avoid extrapolating treatment effects for treated individuals whose baseline profiles are not represented among valid untreated comparison observations. As a result, not every observed enrollment cohort contributes to every event time or outcome. The estimand is therefore the CATT in supported cohort-event-time comparisons, not an extrapolated effect for unsupported regions of the covariate space.

Balance is evaluated within the actual comparison samples used for estimation. Because support restrictions and estimability are applied separately within each outcome-specific cohort-event-time comparison, balance is reported by outcome. This does not imply that different covariates are used by outcome; it reflects that the effective comparison cells can differ across outcomes.

## **D.7 Inference and interpretation**

Uncertainty is reported using standard errors and confidence intervals that account for repeated observations within individuals. Dynamic average effects are interpreted as benchmark event-time averages. Heterogeneity summaries are interpreted as descriptions of the estimated CATT function within supported regions of the data.

Several interpretation limits follow directly from the design. First, the method identifies CATTs under conditional parallel trends, not treatment effects under conditional random assignment in outcome levels. Second, the forest estimates heterogeneity in baseline-differenced outcomes  $z_{i,c,\ell}$ , not heterogeneity in raw outcome levels. Third, the estimated CATTs are predictions for treated individuals in supported comparisons; they should not be interpreted as precise individual-level treatment rules. Fourth, variable-importance rankings indicate which baseline variables help the forest explain treatment-effect variation, but they do not show that changing those variables would change the treatment effect.

The empirical value of the DiD-CF approach in this setting is therefore not that it turns an observational study into a randomized experiment. Its value is that it combines a staggered DiD comparison with flexible, pre-treatment covariate-based summaries of how estimated effects vary across patient profiles. This is the relevant object for assessing whether average program effects conceal economically important differences across enrolled patients.

## References

- Athey, S., Imbens, G., 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America* 113 (27), 7353–7360. <https://doi.org/10.1073/pnas.1510489113>.
- Athey, S., Tibshirani, J., Wager, S., 2019. Generalized random forests. *Ann. Statist.* 47 (2). <https://doi.org/10.1214/18-AOS1709>.
- Bailey, M.J., Goodman-Bacon, A., 2015. The War on Poverty's Experiment in Public Medicine: Community Health Centers and the Mortality of Older Americans. *American Economic Review* 105 (3), 1067–1104. <https://doi.org/10.1257/aer.20120070>.
- Bancalari, A., Bernal, P., Celhay, P., Martinez, S., Sánchez, M.D., 2026. Community health teams and health utilization in El Salvador. *Journal of Health Economics* 106, 103110. <https://doi.org/10.1016/j.jhealeco.2026.103110>.
- Basu, A., Grieve, R., Pritchard, D., Stevens, W., 2019. One size does not always fit all in value assessment. *The American journal of managed care* 25 (11), 540–542.
- Basu, A., Jena, A.B., Goldman, D.P., Philipson, T.J., Dubois, R., 2014. Heterogeneity in action: the role of passive personalization in comparative effectiveness research. *Health Econ.* 23 (3), 359–373. <https://doi.org/10.1002/hec.2996>.
- Basu, A., Meltzer, D., 2007. Value of information on preference heterogeneity and individualized care. *Med. Decis. Making* 27 (2), 112–127. <https://doi.org/10.1177/0272989X06297393>.
- Baxter, S., Johnson, M., Chambers, D., Sutton, A., Goyder, E., Booth, A., 2018. The effects of integrated care: a systematic review of UK and international evidence. *BMC Health Serv. Res.* 18 (1), 350. <https://doi.org/10.1186/s12913-018-3161-3>.
- Borusyak, K., Jaravel, X., Spiess, J., 2024. Revisiting Event-Study Designs: Robust and Efficient Estimation. *Review of Economic Studies* 91 (6), 3253–3285. <https://doi.org/10.1093/restud/rdae007>.
- Callaway, B., Sant'Anna, P.H., 2021. Difference-in-Differences with multiple time periods. *Journal of Econometrics* 225 (2), 200–230. <https://doi.org/10.1016/j.jeconom.2020.12.001>.
- Chaisemartin, C. de, D'Haultfœuille, X., 2020. Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review* 110 (9), 2964–2996. <https://doi.org/10.1257/aer.20181169>.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21 (1), C1-C68. <https://doi.org/10.1111/ectj.12097>.
- Denteh, A., Liebert, H., 2025. Who increases emergency department use? New insights from the Oregon health insurance experiment. *J R Stat Soc Ser A Stat Soc.* <https://doi.org/10.1093/jrssa/qnaf114>.
- Espinoza, M.A., Manca, A., Claxton, K., Sculpher, M.J., 2014. The value of heterogeneity for cost-effectiveness subgroup analysis: conceptual framework and application. *Med. Decis. Making* 34 (8), 951–964. <https://doi.org/10.1177/0272989X14538705>.
- Gavrilova, E., Langørgen, A., Zoutman, F.T., 2025. Difference-in-Difference Causal Forests With an Application to Payroll Tax Incidence in Norway. *J of Applied Econometrics* 40 (7), 727–740. <https://doi.org/10.1002/jae.70001>.
- Glynn, D., Giardina, J., Hatamyar, J., Pandya, A., Soares, M., Kreif, N., 2024. Integrating decision modeling and machine learning to inform treatment stratification. *Health Econ.* 33 (8), 1772–1792. <https://doi.org/10.1002/hec.4834>.

- Goldzahl, L., Stokes, J., Sutton, M., 2022. The effects of multi-disciplinary integrated care on healthcare utilization: Evidence from a natural experiment in the UK. *Health Econ.* 31 (10), 2142–2169. <https://doi.org/10.1002/hec.4561>.
- Golubinski, V., Wild, E.-M., Winter, V., Schreyögg, J., 2020. Once is rarely enough: can social prescribing facilitate adherence to non-clinical community and voluntary sector health services? Empirical evidence from Germany. *BMC Public Health* 20 (1), 1827. <https://doi.org/10.1186/s12889-020-09927-4>.
- Goodman-Bacon, A., 2021. Difference-in-differences with variation in treatment timing. *Journal of Econometrics* 225 (2), 254–277. <https://doi.org/10.1016/j.jeconom.2021.03.014>.
- Grutters, J.P.C., Sculpher, M., Briggs, A.H., Severens, J.L., Candel, M.J., Stahl, J.E., Ruyscher, D. de, Boer, A., Ramaekers, B.L.T., Joore, M.A., 2013. Acknowledging patient heterogeneity in economic evaluation : a systematic literature review. *Pharmacoeconomics* 31 (2), 111–123. <https://doi.org/10.1007/s40273-012-0015-4>.
- Imbens, G., Rubin, D.B., 2015. Causal inference for statistics, social and biomedical sciences: An introduction. Cambridge University Press, Cambridge, New York, NY, Port Melbourne, New Delhi, Singapore, 625 pp.
- Kent, D.M., Paulus, J.K., van Klaveren, D., D'Agostino, R., Goodman, S., Hayward, R., Ioannidis, J.P.A., Patrick-Lake, B., Morton, S., Pencina, M., Raman, G., Ross, J.S., Selker, H.P., Varadhan, R., Vickers, A., Wong, J.B., Steyerberg, E.W., 2020. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. *Ann. Intern. Med.* 172 (1), 35–45. <https://doi.org/10.7326/M18-3667>.
- Khatri, R., Endalamaw, A., Erku, D., Wolka, E., Nigatu, F., Zewdie, A., Assefa, Y., 2023a. Continuity and care coordination of primary health care: a scoping review. *BMC Health Serv Res* 23 (1), 750. <https://doi.org/10.1186/s12913-023-09718-8>.
- Khatri, R.B., Wolka, E., Nigatu, F., Zewdie, A., Erku, D., Endalamaw, A., Assefa, Y., 2023b. People-centred primary health care: a scoping review. *BMC Prim. Care* 24 (1), 236. <https://doi.org/10.1186/s12875-023-02194-3>.
- Krämer, J., Schreyögg, J., Busse, R., 2019. Classification of hospital admissions into emergency and elective care: a machine learning approach. *Health care management science* 22 (1), 85–105. <https://doi.org/10.1007/s10729-017-9423-5>.
- Kreif, N., DiazOrdaz, K., Moreno-Serra, R., Mirelman, A., Hidayat, T., Suhrcke, M., 2022. Estimating heterogeneous policy impacts using causal machine learning: a case study of health insurance reform in Indonesia. *Health Serv Outcomes Res Method* 22 (2), 192–227. <https://doi.org/10.1007/s10742-021-00259-3>.
- Lippi Bruni, M., Ugolini, C., Verzulli, R., Leucci, A.C., 2023. The impact of Community Health Centers on inappropriate use of emergency services. *Health Econ.* 32 (2), 375–394. <https://doi.org/10.1002/hec.4625>.
- Macinko, J., Dourado, I., Aquino, R., Bonolo, P.d.F., Lima-Costa, M.F., Medina, M.G., Mota, E., Oliveira, V.B. de, Turci, M.A., 2010. Major expansion of primary care in Brazil linked to decline in unnecessary hospitalization. *Health affairs (Project Hope)* 29 (12), 2149–2160. <https://doi.org/10.1377/hlthaff.2010.0251>.
- Marafino, B.J., Schuler, A., Liu, V.X., Escobar, G.J., Baiocchi, M., 2020. Predicting preventable hospital readmissions with causal machine learning. *Health Serv. Res.* 55 (6), 993–1002. <https://doi.org/10.1111/1475-6773.13586>.
- OECD, 2023. Integrating Care to Prevent and Manage Chronic Diseases. Best Practices in Public Health. OECD, Paris.
- OECD, 2024. Fiscal Sustainability of Health Systems, 1st ed. Organization for Economic Cooperation & Development, Paris.

- Quinton, J.K., Duru, O.K., Jackson, N., Vasilyev, A., Ross-Degnan, D., O'Shea, D.L., Mangione, C.M., 2021. High-cost high-need patients in Medicaid: segmenting the population eligible for a national complex case management program. *BMC Health Serv. Res.* 21 (1), 1143. <https://doi.org/10.1186/s12913-021-07116-6>.
- Quinton, J.K., Jackson, N., Mangione, C.M., Moin, T., Vasilyev, A., O'Shea, D.L., Duru, O.K., 2023. Differential Impact of a Plan-Led Standardized Complex Care Management Intervention on Subgroups of High-Cost High-Need Medicaid Patients. *Population health management* 26 (2), 100–106. <https://doi.org/10.1089/pop.2022.0271>.
2026. R: The R Project for Statistical Computing. <https://www.r-project.org/> (accessed 30 June 2026).
- Ress, V., Wild, E.-M., 2024. The impact of integrated care on health care utilization and costs in a socially deprived urban area in Germany: A difference-in-differences approach within an event-study framework. *Health economics* 33 (2), 229–247. <https://doi.org/10.1002/hec.4771>.
- Robins, J.M., Rotnitzky, A., Zhao, L.P., 1994. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association* 89 (427), 846. <https://doi.org/10.2307/2290910>.
- Rocks, S., Berntson, D., Gil-Salmerón, A., Kadu, M., Ehrenberg, N., Stein, V., Tsiachristas, A., 2020. Cost and effects of integrated care: a systematic literature review and meta-analysis. *Eur J Health Econ* 21 (8), 1211–1221. <https://doi.org/10.1007/s10198-020-01217-5>.
- ROSENBAUM, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70 (1), 41–55. <https://doi.org/10.1093/biomet/70.1.41>.
- Shah, V., Hatamyar, J., Hidayat, T., Kreif, N., 2026. Exploring the heterogeneous impacts of Indonesia's conditional cash transfer scheme (PKH) on maternal health care utilisation using instrumental causal forests. *Empir Econ* 70 (2), 39. <https://doi.org/10.1007/s00181-025-02846-6>.
- Shah, V., Kreif, N., 2025. Estimating heterogeneous impacts Of subsidised health insurance: A causal machine learning approach. *PloS one* 20 (9), e0315057. <https://doi.org/10.1371/journal.pone.0315057>.
- Sundmacher, L., Fischbach, D., Schuettig, W., Naumann, C., Augustin, U., Faisst, C., 2015. Which hospitalisations are ambulatory care-sensitive, to what degree, and how could the rates be reduced? Results of a group consensus study in Germany. *Health Policy* 119 (11), 1415–1423. <https://doi.org/10.1016/j.healthpol.2015.08.007>.
- Tibshirani, J., Athey, S., Sverdrup, E., Wager, S., 2017. CRAN: Contributed Packages.
- Vuik, S.I., Mayer, E., Darzi, A., 2016. Enhancing risk stratification for use in integrated care: a cluster analysis of high-risk patients in a retrospective cohort study. *BMJ open* 6 (12), e012903. <https://doi.org/10.1136/bmjopen-2016-012903>.
- Wager, S., Athey, S., 2018. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association* 113 (523), 1228–1242. <https://doi.org/10.1080/01621459.2017.1319839>.
- Wammes, J.J.G., van der Wees, P.J., Tanke, M.A.C., Westert, G.P., Jeurissen, P.P.T., 2018. Systematic review of high-cost patients' characteristics and healthcare utilisation. *BMJ open* 8 (9), e023113. <https://doi.org/10.1136/bmjopen-2018-023113>.
- Wild, E.-M., Winter, V., Ress, V., Golubinski, V., 2022. What is the impact of introducing a non-clinical community health advice and navigation service on the demand for primary care in socially deprived areas? Evidence from an observational panel study with difference-in-differences design. *BMJ open* 12 (10), e061964. <https://doi.org/10.1136/bmjopen-2022-061964>.