Università
Ca'Foscari
Venezia

Corso di Dottorato di Ricerca

in Economia

ciclo XXXVI

Tesi di Ricerca

Essays in Causal Inference and Health Economics

SSD: SECS-P/03, SECS-P/05

Coordinatore del Dottorato
ch. prof. Pietro Dino Enrico Dindo

Supervisori
ch. prof. Irene Mammi
ch. prof. Davide Raggi

Dottorando
Andrea Ciaccio
Matricola 956559

Acknowledgments

This dissertation would not have been possible without the help and generous support of many people, both inside and outside academia. A few lines of acknowledgment cannot fully capture the numerous exchanges and insights that have motivated me along this path.

First and foremost, I would like to express my deepest gratitude to my supervisors, Prof. Irene Mammi and Prof. Davide Raggi, for your invaluable guidance, patience, and encouragement throughout this journey. Your expertise and insightful feedback have played a crucial role in shaping this dissertation and my academic growth. I have truly appreciated working alongside you and have been inspired by your dedication and guidance.

I would also like to personally thank Prof. Federico Belotti, Prof. Paolo Li Donni, and Prof. Mario Padula for taking the time to review my thesis and for providing thoughtful and constructive comments and suggestions that have greatly improved the quality of this thesis.

A heartfelt thanks go to Prof. Manolo Arellano, who kindly hosted me during my visiting period at Cemfi. Your support, kindness, and, above all, insightful and constructive discussions made my stay an enriching and inspiring experience. I am grateful for the opportunity to work under your supervision.

Special thanks are also due to Prof. Elisa Tosetti, Prof. Dmitry Arkhangelsky, Prof. Alexander Mayer, and Dr. Domenico Depalo for their insightful discussions and valuable suggestions over these years. In addition, I would like to express my heartfelt gratitude to Prof. Chiara Monfardini for instilling in me a passion for microeconometrics and for always believing in me. I am also grateful to Lisa and Prof. Dindo for their support throughout these years.

What I have learned so far and become the scholar I am is due to you all!

To my PhD colleagues who have shared this journey with me. In particular, I express my deepest gratitude to my cohort mates Angelica and Enrico, for your patience, invaluable support and discussions, and all the laughter and funny moments we have shared over these years. I do not know how I would have survived this journey without your support.

To all the Cemfi PhD students I met during my stay, thank you for your support and hospitality. A special thank you goes to Francesco, Andrea, Paola, Giovanni, and Adriano for making me feel at home, for all the fun we had together, and for your invaluable support. This experience wouldn't have been the same without you!

A sincere thank-you goes to all the friends I have met throughout my academic career. To Simo, Agne, Jack, Ari, and Silvi, for your continuous and immense support, laughter, and wonderful moments. To Annalivia, for your incredible and invaluable support, for listening to me, for your patience, and for all the microeconometrics discussions we have shared over the years. To Francesca, for being a fantastic flatmate and for making the last year of this journey a little less painful.

To all my friends from the South, thank you for your kindness and support over these years. To

GianniDeg, Robert, Gigi, Valentina, and Vallo, I am deeply grateful for your friendship. To my *Biondo*, Anna, Giusy, Chiara, and Nico, for all the crazy things we have done together, the laughter, the late nights, and your incredible patience and support. Thanks to each of you, I wouldn't have become the person I am today.

A heartfelt thank you goes to the two people who have probably shared the most of my path–my best friends. To Pasquale, for being there since the beginning of this journey, the countless times we have laughed together, for all the nights spent talking, for always listening when I needed it most, and, above all, for your patience and unwavering belief in me. To Michele, for your incredible support (and patience, above all), for your loyalty, for always being there, for all the hours spent talking together, and for showing me that I could always count on you. I want to thank you both for truly believing in me–probably more than I have ever believed in myself. None of this would have been possible without your help, and a few lines cannot fully express how grateful I am to have you in my life.

Last but certainly not least, I would like to thank my amazing family. I owe everything I have accomplished to you. To my mom, Paola, and my dad, Michele, for your unconditional love, sacrifices, and unwavering support. To my siblings, Chiara and Francesco, for always being there for me, for your help, and for your unconditional love and support. Thank you! None of this would have been possible without you!

Introduction

Determining causal relationships constitutes a key objective in empirical research across many disciplines. In economics, researchers frequently seek to assess the causal impacts of policy interventions in settings where experimental designs are unfeasible. Evaluating a policy's impact (*ex-post*) is paramount to informing policymakers how successful the policy was in achieving predefined objectives. Ex-post evaluations not only provide critical insights into the efficacy of existing interventions but also inform policymakers about potential adjustments needed to ensure desired outcomes are met. Moreover, they facilitate future interventions' design by leveraging past policy impacts.

A key challenge in policy evaluation is estimating causal effects in non-experimental settings. Unlike experimental studies, where the difference in means estimator between the observed outcomes for the treated and the untreated groups represents a consistent estimator of the average policy impact, observational studies pose additional challenges for the researcher. Specifically, the researcher needs to compare the outcome of interest for the units that are eventually treated/untreated with the one that would have been observed for these units had the policy not been introduced (counterfactual outcomes), which are inherently unobservable (Holland, 1986). The researcher's main challenge lies in finding an appropriate way to recover the missing counterfactual outcomes.

Various methodological approaches exist to estimate causal effects and recover counterfactual outcomes, with their applicability depending on the context of the study. Each method relies on different identifying assumptions, often leading to contrasting results, thus complicating the researcher's task of choosing the most appropriate tool. Nonetheless, selecting the most appropriate causal inference technique is paramount for a sound policy evaluation.

This dissertation addresses two policy-relevant questions in health economics while highlighting the importance of using the most appropriate causal inference tool to evaluate policies' causal effects. In addition, it contributes to the program evaluation literature by developing a novel approach to assess the distributional effect of policies in non-experimental settings characterized by multiple periods and groups.

The first two chapters of this thesis are intertwined and focus on the recent advancements in the *Difference-in-Differences (DiD)* literature in settings involving multiple periods and groups with staggered treatment adoption.

The Difference-in-Differences is one of the oldest and most popular methods employed for estimating policies' causal effects in non-experimental settings. In its classical formulation, we have two periods and two groups: in the first period, no unit is treated; at the beginning of the second period, some units become treated (treated group) while others remain untreated (untreated group). Under the assumption that had the policy not been implemented, the average evolution between treated and untreated units would have been parallel (also known as *common trends* or *parallel trends (PT)* assump-

tion), it is possible to estimate the Average Treatment Effect on the Treated (ATT). This parameter represents the average causal effect of the policy for the treated subpopulation.

The DiD's popularity stems from its broad applicability and the ATT can be consistently estimated using a two-way fixed effects (TWFE) regression if we have many independent clusters. However, many empirical applications involve departures from the canonical two-group, two-period setting.

Recent advancements in the causal inference literature have focused on how departures from the canonical DiD scenario influence the estimation of the ATT (see, for instance, Roth et al., 2023). Unless one is willing to make strong assumptions, it can be proven that the estimated coefficient for the treatment dummy obtained via TWFE represents an inconsistent estimator for the ATT in a context with staggered policy rollout. For this reason, different estimators have been recently proposed to address this issue (Borusyak et al., 2021; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Wooldridge, 2021; De Chaisemartin and d'Haultfoeuille, 2022a).

Building on these advancements, the first chapter of this dissertation assesses the impact of a cost-containment measure first introduced in Italy in 2007 – *Piani di Rientro sanitari* (PdRs) – on the quality and efficiency of Regional Health Services (RHSs). Thus far, ten out of twenty-one RHSs have undergone at least one round of PdRs – three managed to exit, but seven are still treated – raising the question of whether cost reduction has had any unintended adverse effects on the quality of treated RHSs. To answer this question, I exploit the method proposed in Wooldridge (2021), the *Two-way Mundlak approach*. Using Wooldridge's method, which explicitly models the staggered nature of the policy, allows me to analyze how the treatment effect varies across different dimensions. Further, it enables the estimation of the long-run impact of PdRs.

Overall, I find that *Piani di Rientro* managed to reduce costs. However, cost reduction was not followed by a boost in the efficiency of RHSs and the appropriateness of care provided, as expected by the policymaker. Conversely, reduced budgets made available to regions only resulted in an unintended deterioration in the quality of healthcare services. Results also hold in the long run and are robust to other methods that account for the staggered nature of the policy and to a set of *bounded-variations* assumptions (Manski and Pepper, 2018). This study thus contributes to the existing literature documenting that healthcare spending cuts are not necessarily associated with increased efficiency, but can harm health outcomes.

While most studies employ causal inference methods to estimate the ATT, in some cases, policymakers may wish to understand how a policy affects the entire (or specific parts of the) distribution of the outcome of interest. One way to examine the distributional effect of a policy is to consider the Quantile Treatment Effect on the Treated (QTT). The QTT estimates the causal impact of a policy – for the treated group – on a specific quantile of the outcome of interest by comparing the quantiles of treated and untreated outcomes.

If, on one side, different methods have been proposed to estimate QTT in two-period two-group settings (Athey and Imbens, 2006; Bonhomme and Sauder, 2011; Fan and Yu, 2012; Callaway et al., 2018; Callaway and Li, 2019), to the best of my knowledge, no estimator exists to evaluate the distributional impact of the policy in a context with staggered treatment adoption.

To address this gap, in the second chapter, I propose a method to recover the counterfactual quantile and, more generally, the whole distribution of the counterfactual outcome for the treated group in non-experimental settings characterized by multiple groups and periods with variation in treatment timing.

To identify and estimate the full counterfactual distribution, I exploit the intuition behind the es-

timator of the group-time average treatment effects proposed by Callaway and Sant'Anna (2021). By applying to each pair of treated cohorts and never-treated units the method proposed by Callaway and Li (2019), I show that identification and estimation of the entire counterfactual distribution can be achieved. Specifically, identification is reached under a distributional version of the classical Parallel Trends assumption and a new assumption regarding the dependence (or copula) between the change in untreated potential outcome and the pre-treatment level of untreated potential outcome (*copula invariance* assumption).

Besides the QTT, building on Maasoumi and Wang (2019), I consider different approaches that 'anonymously' summarize the quantiles of the distribution of the outcome of interest (such as tests for stochastic dominance rankings) without relying on rank invariance assumptions.

The finite-sample properties of the estimator proposed are analyzed via different Monte Carlo simulations. Despite being slightly biased for relatively small sample sizes, the proposed method's performance increases substantially when the sample size increases.

The third chapter of this dissertation examines the causal effect of environmental crimes on health outcomes, focusing on the illegal dumping and burning of (toxic) waste. An increasing body of scientific literature shows that uncontrolled and illegal waste disposal poses significant health risks, as it may contaminate groundwater, soil, and air. Thus, severe conditions are created for the population residing in areas affected by such phenomena. Besides, the illegal burning of hazardous waste can directly negatively affect the health of nearby communities (WHO, 2024).

To answer the question of whether being exposed to illegal dumping and burning of (toxic) waste may have detrimental effects on health outcomes, we focus on Campania, a region from the South of Italy, and, more specifically, on a particular area of 90 municipalities between the provinces of Naples and Caserta. In this area, known as the Land of Fires (or, in Italian, *Terra dei Fuochi*), organized crime appears to have managed the disposal of toxic waste since the end of the 1980s.

Unlike the previous chapters, this chapter addresses the causal question of interest relying on a selection on observables argument. Specifically, to understand whether living in the Land of Fires may have adverse health effects, we exploit the *inverse probability weighted regression adjustment (IPWRA)* estimator. This method allows us to estimate the average treatment effect (ATE) by controlling for observed demographics and socio-economic differences that may correlate with treatment status. We show that failing to account for observed differences in socio-demographic characteristics leads to biased results.

Our findings indicate that individuals residing in the Land of Fires die more from cancer than in the rest of the region. Although these results hold for both sexes, females experience larger adverse health effects. Furthermore, these results are consistent across different specifications.

Overall, this dissertation answers two policy-relevant questions in health economics using frontier causal inference techniques, highlighting the importance of using the most appropriate tool for a sound policy evaluation. By building on the recent DiD literature, for instance, I show that neglecting the staggered nature of the policy leads to a downward-biased estimation of the effect of Piani di Rientro. Further, I also show that failing to account for observed differences in socio-economic characteristics would lead to a biased estimation of the effect of the Land of Fires on health outcomes. Lastly, I provide a simple method to construct different distributional causal parameters in non-experimental settings with staggered treatment adoption, filling the gap in the distributional treatment effect with selection on unobservables literature.

Contents

List of Figures

List of Tables

The Impact of a Cost-containment Measure on the Quality of Regional Health Services in Italy: a Parametric and a Non-parametric Approach

Andrea Ciaccio[*]

[*]Department of Economics, Ca' Foscari University of Venice

Abstract

Healthcare spending cuts are commonly undertaken, especially in countries with publicly funded healthcare services, to restore budget balance. Understanding whether cost-containment measures have adverse effects on the population's health status is a key concern for policymakers. This paper provides novel evidence on the impact of a cost-containment measure first introduced in Italy in 2007–*Piani di Rientro sanitari* (PdRs)–on the quality and efficiency of Regional Health Services (RHSs). Thus far, ten out of twenty-one RHSs have undergone at least one round of PdRs–three have exited, while seven remain under treatment–raising the question of whether cost reduction has led to any unintended negative effects on the quality of treated RHSs. I answer this question using the *Two-way Mundlak* approach. Compared to the classic Two-way Fixed Effects, this method explicitly accounts for the staggered nature of the policy, allowing me to analyze how the treatment effect varies across different dimensions. Furthermore, it enables the estimation of the long-run impact of PdRs. Overall, I find that *Piani di Rientro* successfully reduced costs. However, cost reduction was not accompanied by improvements in RHS efficiency or the appropriateness of care provided, contrary to policymakers' expectations. On the contrary, the reduced budgets allocated to regions resulted in an unintended deterioration in the quality of healthcare services. These results persist in the long run and remain robust under a set of *bounded-variation* assumptions.

JEL codes: C14, C21, I10, I18, J38

Keywords: Recovery plans, Health outcomes, Variation in treatment timing, Treatment effect heterogeneity, Two-way Mundlak, Bounds

Acknowledgments: I am grateful to my supervisors, Irene Mammi and Davide Raggi, as well as to Federico Belotti, Paolo Li Donni, and Mario Padula for their invaluable advice, insights, and detailed feedback on previous drafts of this paper. I also thank Domenico Depalo for his thoughtful comments on an earlier version. Additionally, I appreciate the valuable feedback from participants at the Ca' Foscari Internal Departmental Seminar, the 11th SIdE Workshop for Ph.D. Students in Econometrics and Empirical Economics, the XXXV SIEP Conference, the XXVIII AIES Conference, and the Ca' Foscari – Groningen Ph.D. Exchange Seminars.

1.1 Introduction

The rising concern about the sustainability of public finances, particularly the need to maintain budget balances, has gathered significant attention from policymakers worldwide, especially following the 2007-2008 financial crisis. In response, some countries have implemented various measures to cut public expenditures, particularly healthcare, drastically. Rationalizing healthcare spending is commonly undertaken to restore budget balance, especially in countries with publicly funded health services. This is mainly due to the substantial portion of GDP that many countries allocate to healthcare. For example, before the COVID-19 pandemic, healthcare expenditure in OECD countries accounted for an average of 8.8% of their GDP.

While healthcare spending is often viewed as a primary candidate for reducing public expenditure, the growing demand for healthcare services and the need for high-quality care have increased concern about whether budget cuts to healthcare systems may negatively affect public health. Numerous studies have investigated whether financial crises and healthcare budget cuts could have adverse effects on population health outcomes (e.g., Crémieux et al., 1999; Quaglio et al., 2013; Franklin et al., 2017; Golinelli et al., 2017; Arcà et al., 2020), as higher health expenditure is usually associated with better health outcomes (OECD, 2021).

The health community has not yet reached a globally accepted definition of quality care (WHO, 2006, 2018). A widely used definition, however, is provided by the United States Institute of Medicine, which defines quality care as "the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge" (Institute of Medicine, 2001). Furthermore, the Institute of Medicine identifies effectiveness, efficiency, equity, patient-centeredness, safety, and timeliness as key dimensions of care. For the remainder of this paper, *quality* will refer to any health service contributing to improved health outcomes.

In addition to providing high-quality services, healthcare systems must also be efficient. This is particularly crucial in resource-limited settings, where improving efficiency is a key priority for policymakers facing increasing demand for health services.

While the health literature lacks a universally accepted definition of quality, the concept of efficiency is even more challenging. According to economic theory, efficiency typically refers to the relationship between inputs (e.g., hospital supply) and either intermediate (e.g., waiting time) or final outputs (e.g., lives saved, quality-adjusted life years) (Palmer and Torgerson, 1999). However, for patients, outcomes matter more than outputs. Over the years, two broad categorizations of efficiency have been proposed. The first, *technical efficiency*, focuses on maximizing outputs or outcomes while keeping input levels constant. The second, *allocative efficiency*, concerns the optimal allocation of resources to achieve the best health outcomes at the lowest possible cost (OECD, 2016). For the remainder of this paper, *efficiency* will refer to allocative efficiency, as the objective is to examine whether budget cuts to healthcare have led to unintended consequences for health outcomes.

Italy represents an interesting scenario for assessing whether healthcare spending cuts may have unintended consequences on health outcomes. The country closely aligns with the OECD average, allocating approximately 8.7% of its GDP to healthcare (OECD, 2021). Despite the Italian National Health Service (NHS) operating as a multi-tier system with substantial regional variability in service quality–since each region manages its own Regional Health Service (RHS)–the introduction at the beginning of the 2000s of Essential Levels of Assistance (ELAs) ensures comparability of healthcare provision across

regions (Aimone Gigio et al., 2018).

Although the decentralization process initiated in the 1990s gradually transferred greater responsibility for RHS management to regional governments to contain costs, the Central Government has continued to finance ex-post the large deficits run by regions. As a result, public health expenditure–and consequently, the overall public health deficit–has increased significantly. To counteract the increased spending and restore financial stability, *Piani di Rientro sanitari* (in English *Recovery Plans (RPs)*) were introduced since 2007. This policy had two main objectives: first, to restore budget balance, and second, to ensure (or improve) ELAs.

There is broad consensus that introducing PdRs successfully reduced costs; however, its impact on quality and efficiency remains mixed. On the one hand, treated regions report declining service quality following the plan's implementation (Calabrò, 2016). On the other hand, the Central Government observes efficiency improvements with no reported deterioration in RHS quality. Nonetheless, some treated regions continue to underperform in terms of ELAs (SiVeAS, 2014; Aimone Gigio et al., 2018). Similar contrasting evidence is found in the literature on RPs (e.g., Depalo, 2019; Bordignon et al., 2020).

This paper provides novel evidence on the causal impact of PdRs on the quality and efficiency of RHSs, contributing to the broader literature on the effects of healthcare budget cuts on health-related outcomes (e.g., Heijink et al., 2013; Borra et al., 2020; Arcà et al., 2020). Specifically, it advances the existing literature in several ways.

First, to the best of my knowledge, this is the first study to assess the long-term impact of the policy. Although two existing studies–Guccio et al. (2023) and Beraldo et al. (2023)–use data up to 2018, as I do in this analysis, they estimate only the overall effect of PdRs, failing to capture the cumulative impact of the policy over time. Assessing long-run effects is essential for a sound policy evaluation. While cost-containment effects may be immediate–since regions undergoing a PdR are required to reduce costs to access additional funding–the structural reorganization of the RHS may take longer to influence service quality and efficiency, as noted by Calabrò (2016).

Moreover, unlike Guccio et al. (2023) and Beraldo et al. (2023), who focus primarily on mortality rates and interregional patient mobility, respectively, this paper offers a more comprehensive analysis by incorporating additional indicators of quality, efficiency, and costs.

Second, this paper contributes to the ongoing debate on whether Piani di Rientro successfully ensured ELAs. Specifically, it expands the analysis by considering a broader set of quality indicators. In addition to evaluating variables directly included in the Ministry of Health's (MoH) ex-post monitoring, this study also examines the impact of PdRs on other indicators that may have been indirectly affected by budget cuts in treated regions. This allows me to test the hypothesis whether regions have strategically outperformed on those indicators that discriminate on their ability to receive funds. If this is the case, following the introduction of RPs, treated regions should exhibit a deterioration in indicators not part of the MoH evaluation process.

A recent strand of the Difference-in-Differences (DiD) literature has highlighted that the estimator for the average treatment effect on the treated (ATT) obtained through the traditional Two-Way Fixed-Effects (TWFE) regression–widely used in previous studies on RPs–can be severely biased in settings with multiple periods and variation in treatment timing, such as the one under analysis (Borusyak et al., 2021; Callaway and Sant'Anna, 2021; Goodman-Bacon, 2021; Sun and Abraham, 2021; De Chaisemartin and d'Haultfoeuille, 2022a,b; Roth et al., 2023). Ignoring the staggered nature of the policy and the possibility that treatment effects vary over time and across regions may thus lead to invalid inference.

To address this issue, the third contribution of this paper is to explicitly account for the staggered nature of the policy when estimating the ATT. This choice is motivated by the DiD decomposition proposed by Goodman-Bacon (2021), which explains why the classic TWFE estimator may be inconsistent for ATT estimation. Additionally, I allow heterogeneity in the treatment effects between regions and over time. The former allows for assessing whether regions that experienced more abrupt cost reductions faced different impacts on the quality and efficiency of RHSs, while the latter enables an evaluation of the policy's long-run effects. To this end, I employ the estimator proposed by Wooldridge (2021), the *Two-Way Mundlak (TWM) approach*.

Although several studies have assessed the impact of Recovery Plans, to the best of my knowledge, this is the first to explicitly model the policy's staggered nature while accounting for treatment effect heterogeneity. Only two papers–Depalo (2019) and Beraldo et al. (2023)–have incorporated heterogeneity to some extent.

Depalo (2019) examines only regions first treated in 2007, evaluating the policy's impact at the end of the first and second rounds of PdRs. As a result, this approach neither accounts for regions that exited treatment over time nor estimates the policy's long-run effects. Beraldo et al. (2023), on the other hand, employs the estimator proposed by Callaway and Sant'Anna (2021) as a robustness check to assess whether the results of their main identification strategy aligned with those of an estimator that accommodates the variation in treatment timing. Using this latter approach, the authors also estimated heterogeneous treatment effects; however, they focused solely on aggregation schemes that assess how the ATT varies with the length of exposure to PdR, without considering heterogeneity across regions. Moreover, unlike Beraldo et al. (2023), this paper considers other indicators beyond the escape rate.

Furthermore, the *subcluster wild bootstrap* proposed by MacKinnon and Webb (2018), which may lead to improved finite-sample inference in a context with few (treated) clusters, will be used.

Lastly, since the consistency of the ATT estimator obtained using the TWM approach relies on three strong assumptions, I build on Depalo (2019) and present the results using the estimator proposed by Manski and Pepper (2018). This estimator allows for assessing the robustness of the results obtained using the parametric estimator by relaxing the main identifying assumptions, as it permits the researcher to directly incorporate the uncertainty about the validity of the 'exact invariance' assumption – exploited to construct the counterfactual outcome (e.g., the parallel trend assumption) – by deriving bounds for the treatment effect. This approach has two key advantages. First, it does not require the researcher to specify the exact reason why the identifying assumption may fail. Second, it enables estimating region-specific treatment effects, offering additional insights into the parametric results.

One of this paper's strengths is performing inference using both methods, as this allows for assessing the results' robustness to the relaxation of the identifying assumptions.

Overall, Piani di Rientro effectively reduced costs. However, contrary to the policymakers' expectations, the policy did not enhance the efficiency of the treated RHSs, except in Abruzzo and Calabria. Moreover, regions that underwent a PdR experienced a deterioration in service quality. These effects persist in the long run and remain robust across various sensitivity analyses. Furthermore, the most significant adverse effects on quality were observed in regions that experienced the sharpest reductions in hospitalization rates. These findings align with existing studies suggesting that healthcare spending cuts do not necessarily lead to efficiency gains but may have detrimental effects on health outcomes (e.g., McKay and Deily, 2008; Siverskog and Henriksson, 2022).

This paper is structured as follows. Section 1.2 outlines the key features of the Italian NHS and the process that led the Central Government to adopt RPs. In Section 1.3, how a PdR works and the government's objectives are exposed. Section 1.4 reviews existing studies on RPs and, more broadly, the health economics literature on cost-containment measures and their impact on health outcomes. In Section 1.5, I describe how the dataset used for the analysis was constructed and the indicators used. Section 1.6 presents the empirical strategy, the possible threats to identification, and different sensitivity exercises. Section 1.7 reports the main findings from both parametric and non-parametric approaches, alongside robustness checks, and discusses these results in light of the existing literature. Finally, Section 1.8 concludes.

1.2 Institutional Setting

The Italian NHS is a regionally-based system founded in 1978 by bailing out the old service built on mutual funds, which was running a large deficit. The system is based on three main guiding principles to comply with the Italian Constitution (*art. 32*): universality, equity, and solidarity. The legislator's primary purpose was to provide uniform services across regions and guarantee equitable access to all citizens, regardless of socio-economic conditions. The system is (almost) fully funded through general taxation.[1] In addition, it is organized into three hierarchical layers to ensure uniform coverage across regions. The Ministry of Health is responsible for developing national health targets to ensure that the general principles are met throughout the country. To reach national goals, regional health departments are entitled to allot spending and to provide benefits packages through Local Health Authorities (LHAs).

Although one of the main objectives of having a decentralized system was to make regional governments more accountable for public health spending and thus contain costs, healthcare expenditure began rising dramatically over the 1980s and at the beginning of the 1990s. In these years, in some regions, health-related spending exceeded the total amount of funds the region had received from the State. To counteract the increased spending, starting in the 1990s, the Central Government initiated a process of budgetary decentralization to settle the national deficit to comply with the Maastricht Treaty's fiscal rules. Different reforms were then implemented to devolve more responsibilities, in managing the Regional Health Systems (RHSs), to the local governments and to boost efficiency.

This decentralization process culminated with the *Riforma del Titolo V* in 2001, which formally attributed to regions the health protection and the responsibility to respect budget balance.[2] Moreover, to ensure each citizen was guaranteed a minimum level of health services and that these were comparable across regions, Essential Levels of Assistance were established at the beginning of the 2000s.[3]

In the years following 2001, however, the Central Government continued to finance, ex-post, the large deficit run by local governments. This fact, coupled with the lack of an adequate mechanism to limit costs by conditioning access to funds, has caused health-related expenditure to grow dramatically over time.

[1] Citizens who are not exempted are required to pay a (light) form of co-payment – *ticket* – whose amount is proportional to the type of service provided.

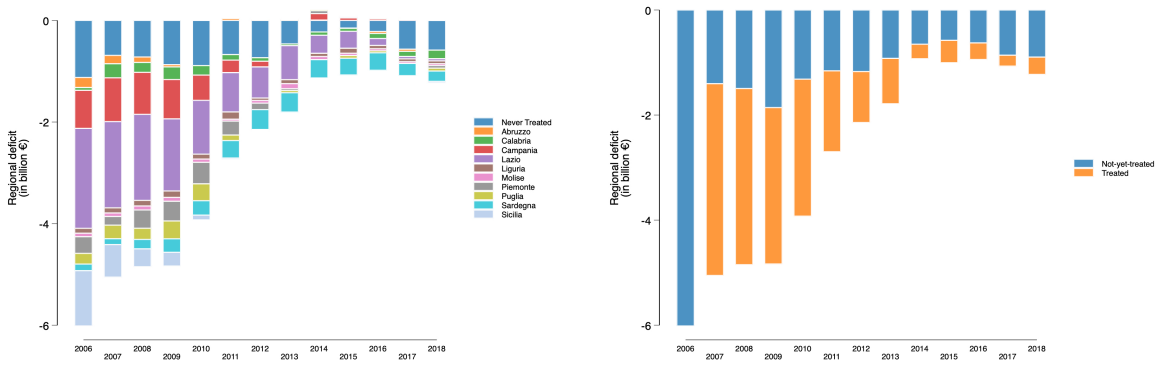
[2] Nowadays, funds are collected mainly through general taxation and redistributed to regional jurisdictions according to the population size and the related distribution by age and gender.

[3] For a thorough review on ELAs, please refer to Torbica and Fattore (2005) and to the Italian Ministry of Health website: https://www.portaletrasparenzaservizisanitari.it/en/prestazionegarantitassn/prestazioni-garantite-dal-ssn.

To curb the excessive spending run by local governments, the State introduced, with the budget law for 2005, a set of measures aimed at constraining access to a part of the health-related resources. Entry was made conditional on evaluating the region's compliance with the budget balance. In case of excessive deficit, to gain access to additional funding, the region was forced to identify the imbalance's potential causes and design a three-year operational plan to restore financial stability through a structural reorganization of the RHS.[4]

Nevertheless, these operational plans have never become effective, and the total health expenditure deficit reached six billion euros only in 2006 (see Figure 1.1). In this context, the Central Government made effective, with the budget law for 2007, the *Piani di Rientro (PdRs) sanitari*.[5]

Figure 1.1: Regional Deficit for Public Health Expenditure



(a) (b)

Notes: These figures report the regional deficit for public health expenditure. Specifically, Panel (a) compares the total regional deficit run by never-treated regions against that run by regions that, at some point, underwent a PdR. Panel (b), instead, depicts the evolution of the regional deficit by comparing regions that were not-yet-treated against those that were treated in a given year.

1.3 Recovery Plans

Starting in 2007, regions with a deficit exceeding 5% of total funding (initially set at 7%) were required to sign region-specific Recovery Plans to access financing.

An RP is a bilateral agreement between the region and the Central Government (specifically, the Ministry of Health and the Treasury), in which the region must identify–through a SWOT analysis (strengths, weaknesses, opportunities, and threats)–the necessary measures to restore budget balance while ensuring (or improving) Essential Levels of Assistance, which represents the RP's *general targets*. To achieve these, RPs also define *specific targets*, outlining intervention areas, and *operational targets*, detailing implementation strategies. In addition to structural reorganization measures within the RHS, regions under an RP must also increase regional tax rates (*IRAP* and *IRPEF*) to increase revenues.

An RP lasts three years but can be renewed for another triennium (*round*) if the region fails to meet its objectives. Furthermore, in cases of persistent deficit or significant failure to meet targets,

[4]For an in-depth review of the institutional background and all the measured introduced starting from 2005, please refer to Aimone Gigio et al. (2018), Depalo (2019), and Bordignon et al. (2020).

[5]Henceforth, I will indiscriminately use the words Recovery Plans, plans, RPs, PdRs, Piani di Rientro, treatment, and policy.

a commissioner is automatically appointed.[6] In such cases, stricter constraints on RHS tax rates and structural resources are imposed.

To ensure the plan's goals are reached, the Central Government provides active support to regions undergoing an RP in terms of monetary resources and monitoring of the RHS. Regarding financial resources, the State makes available additional funds for treated regions to prevent an abrupt decline in health-related spending.[7] The MoH, in collaboration with the Ministry of Economy and Finance (MEF), oversees the management and evaluation of RHSs through the SiVeAS monitoring system. This system consists of an *ex ante monitoring*–where the Central Government ensures that all necessary legal and financial recovery measures are adequately outlined in the plan–and an *ex-post monitoring*–where the MoH conducts quarterly assessments to verify whether budget balance is being restored and ELAs are being maintained. As highlighted by Depalo (2019), a key feature of this policy is its centralized approach, which ensures cross-regional comparability in target achievement.

This policy's primary objectives are to reduce management costs and enhance the efficiency and appropriateness of RHS services. To achieve these goals, regions under a PdR were required to lower hospitalization rates and decrease the number of hospital beds per 1,000 inhabitants. The reduction in hospital beds aimed to encourage a shift toward more cost-effective and appropriate forms of care. In contrast, following a decrease in hospitalization rates, an increase in the percentage of hospital beds occupied and a transition from ordinary to day surgery were expected. While similar measures were introduced nationwide through the 2007 budget law, regions under an RP faced stricter reductions due to their initially higher structural capacity.

In addition to the mandated reduction in structural capacity under PdRs, the 2007 budget law introduced regulations applicable to all regions to control and, ultimately, curb personnel expenses. This provision was subsequently extended through later budget laws in the following years.

Although individual PdRs often imposed further restrictions on new hiring, the automatic hiring freeze was formally introduced only with the 2010 budget law. This measure was applied exclusively to regions under a PdR and only in cases where the region had failed, either partially or entirely, to meet RP targets.

Since 2007, ten out of twenty regions have undergone at least one round of PdRs. The first RP was signed by Lazio in February 2007, followed by Abruzzo, Campania, Liguria, Molise, Sardegna, and Sicilia later that year. Calabria, Piemonte, and Puglia joined in subsequent years (December 2009, July 2010, and November 2010, respectively). Liguria and Piemonte exited the RP after meeting PdR's targets in April 2010 and March 2017, respectively, while Sardegna withdrew in December 2010, invoking its special statute.

Separately, starting in 2008, some regions were placed under the administration of an external commissioner. These include Lazio (2008-2020), Abruzzo (2008-2016), and Campania (2009-2020). Today, the only two regions still under a commissioner are Calabria (since 2010) and Molise (since 2009). A complete timeline of RP adoption in Italy is provided in Table 1.1, which shows the situation as of September 1^{st} of each year.[8]

As shown in Figure 1.1, the Recovery Plans effectively curbed excessive health-related spending. The total deficit declined from 6 billion euros in 2006 to less than 2 billion euros in 2018. However,

[6] Additionally, the Central Government may assign up to two sub-commissioners to oversee plan execution.
[7] Please refer to Depalo (2019) for an overview of how these funds were made available.
[8] Results remain consistent when considering the situation as of June 1st. These are available upon request.

whether cost containment led to efficiency improvements and whether ELAs have been sustained remains unclear. Initially, Government monitoring under the first round of RPs primarily focused on budget balance, but from the second round onward, the Central Government placed equal emphasis on ELAs and financial stability.

Table 1.1: Evolution of RPs in Italian regions (2007–2021)

Region	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Abruzzo	RP	RP	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RP	RP	RP	RP	RP
Calabria				RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC
Campania	RP	RP	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RP
Lazio	RP	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RP
Liguria	RP	RP	RP												
Molise	RP	RP	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC	RPC
Piemonte				RP	RP	RP	RP	RP	RP	RP					
Puglia					RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP
Sardegna	RP	RP	RP	RP											
Sicilia	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP	RP

Notes: This table summarizes the complete history of RPs' adoption in Italy. For each year, the situation on September 1^{st} is reported. Specifically, **RP** denotes if the region is under PdR in year t; while **RPC** if the region at period t is under PdR and the presence of an external commissioner.

1.4 Related Literature

Various studies have examined the impact of Recovery Plans on cost containment and the quality of RHSs. The general consensus is that these plans have effectively reduced costs (Aimone Gigio et al., 2018; Depalo, 2019; Arcà et al., 2020; Bordignon et al., 2020). However, findings on their impact on efficiency and service quality remain mixed, largely due to differences in model specifications and identification strategies across studies.[9]

Exploiting a Two-Way Fixed Effects estimator combined with an instrumental variable approach–using the percentage of citizens complaining about waste in municipal streets as an instrument to address potential violations of the parallel trends assumption–Bordignon et al. (2020) find no statistically significant effect of Recovery Plans on patients' health outcomes. Their analysis incorporates both indicators that are part of the ex-post monitoring carried out by the Central Government and variables that, instead, are not part of the monitoring (such as the mortality rate) as proxies for quality.

In contrast, Cirulli and Marini (2023), also employing a TWFE estimator but using the percentage of citizens dissatisfied with public transport as an instrument, report a significant increase in the total mortality rate, suicide rate, mortality rates due to cancer and heart disease, and discharge rates for psychological disorders in regions affected by RPs.[10]

Similarly, Guccio et al. (2023), applying a Matching Method for Time-Series Cross-Sectional Panel coupled with a TWFE estimator, find a negative impact of RPs on total mortality rates. Likewise, Beraldo et al. (2023), using a TWFE estimator, highlight unintended adverse effects on the escape rate following the introduction of RPs.

[9]Notably, all existing studies on RPs rely on publicly available data–sourced from the Health For All database or the Ministry of Health–which are aggregated at the regional level. As a result, the unit of observation is RHS r at time t.

[10]It is worth noting that both studies use similar sample periods: while Bordignon et al. (2020) analyze data from 2000 to 2014, Cirulli and Marini (2023) extend the period by only two years (1999-2015).

Finally, Depalo (2019), employing a non-parametric approach, identifies a sudden drop in the hospitalization rate within treated RHSs, followed by a rise in all-cause mortality by the end of the first round of RPs. The decline in hospitalization rates is larger in regions experiencing more significant reductions in health spending, and the greater the drop in hospitalizations, the higher the mortality.

Regarding efficiency, Depalo (2019) finds that the efficiency of RHSs–measured by the ratio between the hospitalization for the diagnosis-related group (DRG) at high risk of inappropriateness versus those that could not be avoided–did not improve in treated regions by the end of the first round of Recovery Plans (RPs).

Similarly, using Data Envelopment Analysis (DEA), Giancotti et al. (2020) report no impact of RPs on hospital efficiency between 2010 and 2013. Their analysis relies on a technical efficiency index, constructed as an output-to-input ratio, where hospital bed capacity and hospital staff serve as inputs, and the total number of inpatient discharges represents the output.

Conversely, Guccio et al. (2022), applying a similar methodology to Giancotti et al. (2020) but using data from 2003 to 2010, find a rise in inefficiency in treated regions following the implementation of RPs. Their analysis estimates cost inefficiency using Stochastic Frontier Analysis (SFA), suggesting that, rather than improving efficiency, RPs may have exacerbated inefficiencies in healthcare spending.

All the studies mentioned so far belong to the larger international literature on health economics, which examines the effects of healthcare spending cuts on both efficiency and quality of healthcare providers.

Numerous papers have analyzed whether healthcare expenditure affects health outcomes. Crémieux et al. (1999) investigates whether healthcare spending affects health outcomes using data for Canada over 15 years. The study finds that following a reduction in healthcare spending, a statistically significant increase in infant mortality and a decrease in life expectancy were observed. These results remain robust after controlling for various factors. Using fixed effects and a panel from 1980 to 1995 for 15 EU members, Nixon and Ulmann (2006) documents statistically significant improvements in infant mortality and marginal improvements in life expectancy following increased healthcare spending.

Similar results are found for other proxies of quality, such as mortality. A large body of literature examines the relationship between healthcare spending and mortality. Using a panel of 14 Western countries from 1996 to 2006, Heijink et al. (2013) evaluates the relationship between healthcare spending and avoidable mortality. The authors report that countries with above-average expenditure growth experienced a more significant decrease in avoidable deaths. Similarly, using data from 2005-2006 on primary care trusts in the UK and employing a Generalized Method of Moments estimation procedure with several instruments, Andrews et al. (2017) shows that mortality (from all causes) is highly sensitive to variations in health spending. In particular, they estimate an elasticity of -0.705 for the mortality rate in response to changes in healthcare expenditure. Additionally, employing data from the Spanish Ministry of Health (1996-2015) and Ruhm's Fixed Effects model, Borra et al. (2020) finds a significant increase in mortality from circulatory diseases and external causes following a reduction in hospital supply. Similarly, Siverskog and Henriksson (2022), using 2001-2019 data for Swedish regions and a TWFE model, analyzes the impact of reduced bed capacity on health outcomes. Despite a nationwide decline in bed capacity and mortality during the period, the most significant reduction in mortality rates occurred in regions where the decrease in hospital beds was less pronounced. Moreover, the authors report that providing an additional bed generates three additional quality-adjusted life years (QALYs), supporting the hypothesis that bed capacity is directly linked to patient outcomes.

Further studies document a positive relationship between healthcare spending and efficiency. For instance, using U.S. data and a two-stage semi-parametric DEA, Hunt and Link (2020) analyzes the relationship between public health spending and the technical efficiency of surrounding hospitals. The study finds a positive relationship between areas with high per capita public health expenditure and efficiency gains. McKay and Deily (2008), using U.S. data from 1999-2001 and SFA, examines the relationship between cost inefficiency and health outcomes (proxied by mortality and complication rates). While they fail to find a statistically significant association between cost inefficiency and health outcomes, this result varies across geographical areas. They also emphasize the importance of distinguishing between costs related to efficiency and those due to waste or other inefficient uses of resources. Specifically, reducing cost inefficiency is unlikely to negatively affect health outcomes, while generalized cost reduction may harm health.

Similar findings regarding healthcare spending, health outcomes, and efficiency have been observed in Italy. For example, using a 15-year panel of Italian RHSs, Golinelli et al. (2017) investigates the relationship between per capita healthcare expenditure and total mortality. Overall, they find that reducing healthcare spending is associated with an increase in the mortality rate, holding other factors constant.

Similarly, using a 10-year (2004-2014) panel of Italian regions and exploiting the introduction of PdRs as an instrument to address potential endogeneity issues in healthcare spending, Arcà et al. (2020) estimates a 4.5% increase in avoidable mortality following a €100 reduction in per capita health expenditure. This increase in avoidable deaths appears to result from a drastic reduction in hospital supply (beds and staff) and decreased hospitalization rates. Furthermore, the authors document that patients' mobility flows–mainly from Southern regions to the Center-North of Italy–increase significantly following a decline in healthcare spending, likely due to reduced hospital supply in Southern regions after introducing PdRs.

Inter-regional patient mobility is strongly affected by hospital supply. Using a dynamic spatial panel data model and Italian hospital discharge records for 2001-2010, Balia et al. (2018) discovers that, ceteris paribus, a larger number of hospital beds in the region of origin reduces the likelihood of seeking care in another region. At the same time, the outflow toward a given region increases with hospital supply. While excess hospital beds may indicate poor management, resource waste, and low service quality, larger bed capacity can also indicate lower waiting lists.

Ghislandi et al. (2025) examines the effects of cost-containment measures introduced in Italy after the financial crisis. Using the estimator proposed by Callaway and Sant'Anna (2021) and Italian data from 2008 to 2015, they find that hospital closures negatively impact health outcomes, proxied by the probability of in-hospital mortality and length of stay after acute myocardial infarction. These effects are mediated by increased travel time due to closures.

Regarding the impact of efficiency on health outcomes, Martini et al. (2014) employs 2008-2011 data on hospitals at the ward level in Lombardia and a three-step estimation procedure to analyze whether the pressure for cost containment influences hospital performance. Overall, they estimate a larger mortality rate in more efficient hospitals, but also find that more efficient hospitals exhibit lower readmission rates.

Overall, existing studies suggest that measures aimed at reducing costs through cuts in healthcare spending and hospital supply may adversely affect patients' health. Moreover, these policies have often proven ineffective in promoting health providers' efficiency. For these reasons, it is crucial to under-

stand whether the introduction of PdRs has had unintended effects on the quality of RHSs.

1.5 Data

1.5.1 Data Sources & Sample Selection

A unique dataset gathering information from different administrative sources is built to assess the impact of Recovery Plans. The percentage of patients (aged 65+) diagnosed with hip fracture operated on within 48 hours in an ordinary regime and the ratio between the hospitalization for the diagnosis-related group (DRG) at high risk of inappropriateness versus those that could not be avoided are made publicly available by the MoH through yearly reports, the *Monitoraggio Griglia Lea* (MoH, 2014, 2015, 2016, 2017, 2018). Data on Recovery Plans were directly obtained from the MoH's website.[11] In contrast, all the other indicators of RHSs' quality and costs, as well as controls, were downloaded through the *Health for All (HFA)* software.[12] This latter is a database gathering different aspects of the Italian NHS, made publicly available by the Italian National Institute of Statistics (ISTAT).

Regarding the sample selection, data are already aggregated at the RHS level and include information on all 21 Italian RHSs (the 19 regions and the 2 autonomous provinces of Bozen and Trento). As for the time span, given the availability of a broader pre-treatment period for some of the outcomes of interest, the maximum number of available pre-treatment years for each dependent variable has been exploited to improve the estimators' performance in the following analysis. On the other side, since up-to-date data are unavailable, the analysis is restricted to the end of the fourth round of PdRs, which ended in 2018.[13] The reasons behind this latter choice are twofold. On one side, the analysis is confined to the end of the last completed round for which data are available; on the other side, the effects of the policy do not cumulate to that of the recent Covid-19 pandemic. Overall, at most, 441 region-year observations are available depending on the outcome under analysis.

1.5.2 Indicators

Different proxies for quality, which may directly or indirectly affect population health, are used to assess whether adopting an RP has influenced the quality of RHS. In addition to considering variables directly included in the MoH's ex-post monitoring, the impact of PdRs on other indicators is also evaluated. The policy may have indirectly affected these indicators by reducing the budget available to the region. Specifically, I test the hypothesis that regions strategically outperformed on indicators determining their ability to receive funds. Following the introduction of RPs, treated regions should exhibit a deterioration in indicators not included in the MoH evaluation process if this holds.

The first proxy for RHS quality is the*hospitalization rate* (*total* and *acute*), defined as the number of hospitalizations over the corresponding population multiplied by $1,000$. Hospitalization rates are directly part of the ex-post monitoring carried out by the Central Government (see Section 1.3). They

[11] Please refer to https://www.salute.gov.it/portale/pianiRientro/homePianiRientro.jsp for additional information.

[12] Further details on the choice of the specific indicators and controls used in the following analysis will be given in the next two subsections.

[13] The available last year, to date, is 2019 for most of the dependent variables from the HFA database, 2020 for those from MEF or MoH.

are widely used in health literature as a proxy for healthcare utilization (Depalo, 2019; Arcà et al., 2020) and as an indirect indicator of patients' Quality of Life (Berchialla et al., 2010).

Given the well-documented negative (indirect) impact of healthcare spending cuts on mortality (e.g., Heijink et al., 2013; Andrews et al., 2017; Golinelli et al., 2017; Arcà et al., 2020), the *total mortality rate* is also considered a proxy for the quality of RHS. This rate is measured as the number of deaths (all causes) in a given year divided by the region's population, multiplied by 10,000.

Another quality indicator used in this analysis is the *mortality rate from ischaemic heart diseases*, defined similarly to the total mortality rate.[14] Although not directly included in the ex-post monitoring of the Central Government, numerous studies employ the Acute Myocardial Infarction (AMI) mortality rate–part of this indicator–as a measure of healthcare provider quality (Pross et al., 2017; Schiele et al., 2017). Moreover, despite a substantial decline in coronary heart disease mortality over recent decades, AMI remains a leading cause of death in many OECD countries, including Italy (OECD, 2021). Therefore, assessing whether PdRs influenced this indicator, specifically whether the decline in ischaemic heart disease mortality slowed following the policy's introduction, is crucial.

Free patient choice is a defining feature of the Italian healthcare system, allowing patients to seek care in hospitals outside their home RHS.[15] The Italian NHS exhibits significant inter-regional patient mobility under the ordinary regime–7.2% of total hospitalizations in 2020 alone (MoH, 2020)–with Central and Northern regions acting as net exporters of hospital care (Balia et al., 2018). The substantial share of patients from the South seeking care elsewhere highlights persistent disparities in RHS quality. As "the compensation of net patient flows has generated additional financial resources for central-northern regions, exacerbating the north-south gradient in the Italian NHS" (Balia et al., 2018, p.3), it is crucial to examine the impact of PdRs on the *percentage of patients migrating to other regions for ordinary acute hospitalization*.

The impact of PdRs on the *percentage of patients (aged 65+) diagnosed with hip fracture operated on within 48 hours in an ordinary regime* (hereafter, *% of patients with hip fracture replacement*) is also estimated. This variable is widely used in the health literature as an indicator of the quality of healthcare provider (Pross et al., 2017; OECD, 2021), given the significant decline in quality of life following a hip fracture (e.g., Amarilla-Donoso et al., 2020). Additionally, % of patients with hip fracture replacement serves as a proxy for healthcare system responsiveness. Evidence suggests that timely surgical intervention–within 48 hours–improves patient outcomes and reduces the risk of complications (OECD, 2021), as delays are strongly associated with increased mortality and a higher incidence of pressure sores (Moja et al., 2012).

To assess whether RPs affected RHS efficiency, following Depalo (2019), the *ratio between the hospitalization for the diagnosis-related group (DRG) at high risk of inappropriateness versus those that could not be avoided* (in short, *efficiency ratio*) (hereafter, *efficiency ratio*) is used as a proxy for the inefficient use of healthcare services. Although measuring healthcare provider efficiency is challenging (McGuire, 1987), inappropriate hospitalizations are widely recognized as an indicator of resource misallocation (Angelillo et al., 2000; Navarro et al., 2001; Pileggi et al., 2004). Moreover, the MoH acknowledges this variable as a measure of service appropriateness and efficiency, incorporating it into the Monitoraggio

[14] This indicator includes diseases classified as 410-414 (ICD-9-CM) and 120-125 (ICD-10-CM).

[15] Although patients must cover travel costs, their region of residence reimburses the treatment through a compensation scheme based on DRG tariffs.

Griglia Lea reports.[16]

Another widely used proxy for resource misallocation in health economics is the cesarean sections (c-sections) for first-time mothers (OECD, 2019). While it is well recognized in the medical literature that c-sections can be lifesaving and, at the same time, necessary surgeries, vaginal birth should be preferred as a delivery mode in all such cases when there are no complications or specific reasons (Gregory et al., 2012; OECD, 2019). Two key reasons justify using this indicator as a proxy for inappropriate care. First, although both modes of delivery carry risks (Gregory et al., 2012), c-sections are associated with higher maternal and neonatal mortality, increased risk of complications, and short- (e.g. greater likelihood of allergies and asthma) and long-term (e.g., impaired cognitive development) adverse effects on the child due to the invasiveness of the procedure (Polidano et al., 2017; Sandall et al., 2018; OECD, 2019). Second, c-sections are significantly more costly than vaginal deliveries.

The MoH also recognizes c-sections for first-time mothers as a valid proxy for service appropriateness and efficiency, incorporating it into the Monitoraggio Griglia Lea reports. However, as this indicator is available only for 2012-2016, the $\%$ *c-sections*–defined as the number of c-sections over total childbirths multiplied by 100–will be used instead.[17]

Regarding costs, two indicators–both included in the ex-post monitoring conducted by the Central Government–are employed: *log* of *current health expenditure* and the *number of hospital beds*. The former directly reflects regional healthcare expenditure and aligns with budget balance constraints set by the Maastricht Treaty's rule (Eurostat, 2013). The latter represents a key mechanism for cost containment, as reducing hospital bed capacity is a direct channel of cutting management costs (Aimone Gigio et al., 2018; Arcà et al., 2020).

1.5.3 Socio-economic Characteristics

A set of socio-economic variables that may explain, in the subsequent analysis, differences in the outcome of interests due to observables was obtained through the HFA software. These include region's *population size*, *population distribution by gender*, and *population distribution by age*. As noted in Section 1.2, these factors determine how the central government allocates funds to regional jurisdictions.

Following Depalo (2019) and Balia et al. (2018), *GDP per capita* is also considered to account for income effects. At the micro-level, GDP can serve as a proxy for patients' ability to seek care in the private health sector following a reduced quality of services provided in the public hospital. At the macro-level, it reflects the capacity of wealthier RHSs to provide higher-quality care.

Table 1.2 presents descriptive statistics for these variables, split by treatment cohort. Specifically, it reports means with standard deviations in parentheses. Additionally, columns **Diff.** shows the results of a difference-in-means test comparing the treated cohort r with the never-treated group, with standard errors (se) in parentheses.

On average, treated cohorts are poorer than never-treated RHSs, which aligns with expectations–except for Piemonte, all treated regions are in Central-Southern Italy. Treated cohorts also tend to have larger populations, with a higher share of individuals aged $15-34$, and a smaller share aged $35-54$.

[16]The complete list of DRGs classified as high risk of inappropriateness is provided by the Italian MoH. Examples of such DRGs can be found in Appendix A.

[17]To verify whether these two indicators convey the same information, I regressed the MoH-provided indicator ($\%$ c-sections for first-time mothers) on the HFA-based indicator ($\%$ c-sections), controlling for region and year fixed effects. The resulting R^2 of 0.90 suggests a strong correlation between the two measures.

Interestingly, no clear pattern emerges regarding the percentage of females.

To assess the stability of these characteristics over time, Table 2 in Appendix A provides descriptive statistics split into pre- and post-treatment periods, based on the year the first PdR was signed (2007). Means are reported with standard deviations in parentheses, alongside a differences-in-means test comparing pre- and post-2007 averages, with se in parentheses.

The results show that, apart from the percentage of females and population size–which remained stable over time (except for RHSs first treated in 2011)–GDP per capita increased across all groups, and all groups experienced population aging.

Given the observed differences in baseline characteristics and the fact that most of the variables presented represent the criteria according to which the central government allocates funds, failing to account for these differences could undermine the validity of the analysis.

Table 1.2: Descriptive Statistics Socio-Economic Characteristics

	Never treated	**d2007**	**Diff.**	**d2010**	**Diff.**	**d2011**	**Diff.**
Population size	2499197.76	3009090.39	509892.64*	3153406.52	654208.77*	4070308.71	1571110.96***
	(2734466.99)	(2143484.83)	(252239.16)	(1189634.72)	(257031.67)	(69200.41)	(180547.33)
% people aged 15-34	22.82	24.26	1.43***	24.13	1.31*	26.29	3.46***
	(2.84)	(3.49)	(0.34)	(3.54)	(0.58)	(2.91)	(0.66)
% people aged 35-54	29.83	29.19	-0.64***	28.91	-0.91***	28.44	-1.38***
	(1.28)	(1.47)	(0.15)	(1.27)	(0.21)	(1.17)	(0.27)
% people aged 55-64	12.51	12.29	-0.21	12.41	-0.10	11.80	-0.71**
	(0.82)	(1.20)	(0.11)	(1.25)	(0.20)	(0.91)	(0.21)
% people aged ≥ 65	21.24	20.42	-0.82*	20.74	-0.50	18.16	-3.08***
	(2.40)	(3.60)	(0.34)	(2.66)	(0.44)	(2.19)	(0.50)
% female	51.38	51.60	0.22***	51.40	0.02	51.46	0.09**
	(0.38)	(0.53)	(0.05)	(0.27)	(0.05)	(0.05)	(0.03)
GDP per capita	29312.10	21951.50	-7360.61***	21552.32	-7759.78***	16345.75	-12966.35***
	(6162.61)	(5889.49)	(632.74)	(6408.90)	(1068.81)	(1534.75)	(525.90)

Notes: This table reports descriptive statistics by treated cohort, where means are presented with standard deviations in parentheses. Additionally, it shows differences in means between treated cohort r and the never-treated group, with standard errors (s.e.) in parentheses. **Never-treated** = Regions that have never undergone an RP. **d2007** = Regions that first signed an RP in 2007. **d2010** = Regions that first signed an RP in 2010. **d2011** = Regions that first signed an RP in 2011. ***, **, and * denote significance at 1%, 5%, and 10%, respectively.

1.5.4 Potential Effects & Channels

While some of the indicators described earlier are directly influenced by RPs–either because they represent channels through which the policy operated or because they are part of the ex-post monitoring carried out by the Central Government–others are not explicitly included in the ex-post monitoring framework. However, these indicators may still be affected by cost-containment measures.

Although this study considers multiple indicators, data limitations prevent us from examining how PdRs have affected patient outcomes (e.g., survival after hospital procedures, hospital readmission, etc.). Moreover, the unavailability of more granular data, such as individual- or hospital-level information, hinders a direct investigation into the specific channels through which indicators not explicitly targeted by the policy may have been affected, such as changes in hospital supply and capacity beyond the number of beds (e.g., hiring freezes, and suspension of new staff recruitment).

In fact, while RPs aim to reduce regional healthcare expenditures by reducing inefficient spending components, this often translates into an accounting exercise in which spending is reduced roughly linearly. This can lead to reduced public services or attempts to eliminate redundancies (e.g., staff reductions, and fixed-cost cuts). However, how cost reductions are achieved varies substantially both

within and between RHSs. Unfortunately, the available data do not allow further investigation of this heterogeneity.

Nonetheless, potential channels through which PdRs may have influenced the indicators examined in the following analysis and their expected effects can be inferred from national and international literature. Extensive research has documented that changes in hospital supply and staff levels negatively impact health outcomes (see previous section). A key effect of PdRs was a substantial reduction in NHS personnel, including medical, nursing, administrative, and technical staff (Bordignon et al., 2020). This workforce reduction, implemented by treated regions, not only decreased staff availability, but also led to longer shifts for remaining hospital personnel and increased waiting lists for hospital procedures, potentially affecting quality of care. Another significant consequence of budget cuts implemented by regions under PdRs was the closure of numerous hospitals, particularly in southern Italy. Increased travel time to nearby healthcare facilities has been linked to worse health outcomes, as demonstrated by Ghislandi et al. (2025).

Regarding the impact on hospitalization rates, a reduction in hospitalizations may, on the one hand, suggest that PdRs were effective in cost containment. On the other hand, a decrease in hospitalization rates is generally associated with worse health outcomes (Berchialla et al., 2010; Depalo, 2019; Arcà et al., 2020).

On one hand, PdRs directly affect hospitalization rates; on the other hand, the mortality rate from ischaemic heart diseases is likely to be indirectly impacted by the policy. Three main channels are expected through which PdRs may have affected this indicator. First, the complexity of healthcare management and the high risk of re-hospitalization make AMI, and more generally ischaemic heart diseases, particularly costly for healthcare systems (Kwok et al., 2018; Lobo et al., 2020). Therefore, healthcare spending cuts may have adverse (indirect) effects on these outcomes by reducing the budgets available to the RHS. Second, previous analyses have documented that AMI is sensitive to variations in hospital supply. Increasing hospital beds has been associated with lower 30-day readmission rates after AMI (Brown et al., 2014), while reducing hospital supply has been linked to a significant increase in mortality from AMI (Borra et al., 2020). Third, there is evidence that hospital closures, following the introduction of PdRs, are associated with an increased probability of in-hospital mortality after AMI due to the increased travel distance to nearby hospitals (Ghislandi et al., 2025).

As for the percentage of patients migrating to other regions for ordinary acute hospitalization, if larger outflows are detected in treated regions (mainly from the South of Italy) following the introduction of RPs, this would suggest that the perceived quality in the South has deteriorated and that PdRs have exacerbated the gradient between the South and North of Italy in terms of NHS quality. There are two potential channels through which RPs may indirectly affect patient mobility. First, spending cuts in healthcare expenditure increase the flows from the South to the North of Italy (Arcà et al., 2020). Second, evidence shows that the lower the supply in the region of origin, the higher the likelihood of patients seeking care in another region (e.g., due to lower waiting lists for complex procedures) (Balia et al., 2018).

On the other hand, it is unclear whether PdRs may have adversely affected the percentage of patients receiving hip fracture replacement. On one hand, like AMI, hip fracture repair also constitutes a significant burden for the healthcare system (Williamson et al., 2017). Therefore, healthcare spending cuts will indirectly impact this indicator by reducing the budget available to hospitals and decreasing hospital supply (beds and staff). The timely response from healthcare providers is known to be

influenced by hospitals' operating theatre capacity, flow, and access (OECD, 2021). Conversely, the percentage of patients receiving hip fracture replacement is one of the specific targets outlined in RPs. Thus, regions have strong incentives to monitor this indicator closely and ensure adequate levels are achieved to access financing.

Similarly, it is unclear a priori whether PdRs may have increased or decreased the cesarean rate. On one hand, since this indicator is part of the ex-post monitoring carried out by the Central Government, treated regions have economic incentives to ensure that adequate levels are maintained. Additionally, c-sections are a more costly procedure than vaginal births, which could lead to a reduction in the percentage of c-sections after the introduction of RPs. On the other hand, the introduction of PdRs may have had a positive effect on the cesarean rate. One reason for expecting a positive impact is that, in Italy, cesarean rates are higher in private hospitals than in autonomous public ones (De Luca et al., 2021). If an increase in the percentage of c-sections is observed following the introduction of the policy, this could indicate that the perceived quality of healthcare services in public hospitals has deteriorated, prompting more women to seek care in private hospitals.

Lastly, hospital beds are often used as a proxy for the capital factor in hospital production (Santías et al., 2011; Giancotti et al., 2020; Guccio et al., 2022). While excess staffed hospital beds may signal inefficient resource use, reducing hospital beds can indicate a lower capital factor. Further, numerous studies report adverse health effects following hospital bed reductions (Arcà et al., 2020; Borra et al., 2020; Siverskog and Henriksson, 2022). This is because, holding the number of beds constant, higher bed occupancy rates have been linked to increased mortality (Madsen et al., 2014; Boden et al., 2016), reduced admission rates, and higher readmission rates (see, for example Blom et al., 2014). Thus, although reducing hospital beds may signal the reduction of inefficient spending, it can also result in unintended adverse effects on health outcomes.

1.6 Empirical Strategy

1.6.1 Two-Way Mundlak Approach

As discussed in Section 1.3, Recovery Plans were first introduced in 2007. Although seven regions were required to adopt a PdR that year, the remaining three treated regions implemented a PdR in subsequent years.[18] This framework is called *staggered treatment adoption* setup in the program evaluation literature.

Since this analysis aims to understand whether the introduction of RPs has had any unintended negative consequences on the quality and efficiency of treated RHSs, the causal estimand of interest will be the ATT in periods where the treated regions are effectively under Recovery Plans.

By integrating the dynamic potential outcomes framework (Robins, 1986, 1987) with the dynamic treatment adoption setting (Heckman et al., 2016), we define the ATT for regions first treated in period t and evaluated at time t as follows (see Appendix B for more details):

$$\tau_{r,t} = \mathbb{E}\left(y_{g,t}(r) - y_{g,t}(\infty)|d_r = 1\right) \quad g = 1,\ldots,21, r \in \{q,\ldots,2018\}, t \in \{q,\ldots,2018\} \tag{1.1}$$

[18] For simplicity, the terms Regional Health Service (RHS) and region are used interchangeably in the following sections. However, it is essential to note that Italy has 20 regions, while the number of RHSs is 19, plus the two autonomous provinces of Bozen and Trento.

where $y_{g,t}(r)$ is the potential outcome for region g at time t had the policy been introduced by period r, while $y_{g,t}(\infty)$ denotes the corresponding outcome in period t had the region never been treated. Here, q represents the first period in which the policy is implemented and d_r is a dummy variable that takes value 1 if region g was first treated in period r, and 0 otherwise.

Most existing studies on Recovery Plans rely on variations of the classic Two-Way Fixed Effects regression–such as the one presented in (B.2) in Appendix B–to assess the impact of RPs on health-related outcomes. These studies typically interpret the estimated coefficient for the treatment dummy as an estimator of the ATT.

However, it can be shown that the estimated coefficient for the treatment dummy obtained via a TWFE represents an inconsistent estimator for the ATT in a context with a staggered policy rollout and potentially heterogeneous treatment effects (Borusyak et al., 2021; Goodman-Bacon, 2021; De Chaisemartin and d'Haultfoeuille, 2022a).[19]

To estimate the causal effect of the policy, we will resort to the *Two-Way Mundlak Approach* (Wooldridge, 2021), which provides a consistent estimator of the ATT in contexts like the one under analysis. Specifically, this estimator explicitly models the staggered nature of the policy.

To further support the choice of using an estimator that explicitly accounts for the nature of the policy, I will apply the DiD decomposition proposed by Goodman-Bacon (2021) (see Appendix B for a detailed review). This procedure will help illustrate why the TWFE estimator can be inconsistent in the presence of variation in treatment timing and identify potential sources of bias.

An advantage of the TWM approach over TWFE is that it allows researchers to estimate heterogeneous treatment effects, where heterogeneity may arise between treatment cohorts, treatment intensities, calendar time, and/or covariates (see Appendix B for a detailed review of this method). From a contextual perspective, accounting for heterogeneity across regions and over time is essential for two main reasons.

First, in terms of heterogeneity over time, while the core characteristics of the policy have remained relatively stable, the law formally emphasized both Essential Levels of Assistance and financial stability only from the second round of RPs. In contrast, the first round of PdRs primarily focused on restoring budget balance within RHSs. Additionally, while cost-containment effects may be immediately observable–since regions under a PdR must implement cost-cutting measures to access additional funding–the broader structural reorganization of RHSs takes time. As a result, significant effects on service quality and efficiency may only become evident in the long run, as also noted by Calabrò (2016).

Second, although there is no strong reason to expect systematically different effects across treated cohorts, PdRs are region-specific agreements. Consequently, it is reasonable to expect variations in effects, such as different impacts in Calabria compared to Lazio. This justifies allowing for heterogeneity across regions in the analysis.

For the remainder of this paper, we adopt the notation introduced by Wooldridge (2021). Consider the following regression model:

$$y_{g,t} = \alpha + \sum_{s=-6}^{-2} \tau_s intens_{g,t}(s) + \sum_{s=0}^{11} \tau_s \left(w_{g,t} \cdot intens_{g,t}(s)\right) + \sum_r \lambda_r d_{g,r} + \eta_t + u_{g,t} \tag{1.2}$$

[19] For a review of which are the common assumptions made in the program evaluation literature to estimate the ATT, how the TWFE estimator works, and why the TWFE may be proven to be an inconsistent estimator for the ATT in the such contexts, please refer to Appendix B.

$y_{g,t}$ represents one of the outcomes of interest (see Section 1.5) for region g at time t, where $g = 1, \ldots, 21$, and $t = y_{first}, \ldots, 2018$. $y_{first} = min\{t : y_{g,t} \neq \texttt{NaN} \quad \forall g\}$. That is, y_{first} represents the first year in which the dependent variable of interest is observed (non-missing) for all the RHSs.

Let $w_{g,t}$ denote a binary treatment with support in $\{0, 1\}$. Since the main goal of this paper is to evaluate the impact of RPs, we define $w_{g,t} = 1$ if region g at time t is under an RP, and $w_{g,t} = 0$ otherwise.[20]

η_t represents year fixed effects, and $u_{g,t}$ is the error term. The constant is denoted by α, while $d_{g,r}$'s are mutually-exclusive cohort dummies indicating the year in which region g first received the treatment. Since new regions underwent an RP only in 2007, 2010, and 2011, then $r \in \{never, 2007, 2010, 2011\}$, such that $d_{2007} + d_{2010} + d_{2011} + d_{never} = 1$.

Finally, the term $intens_{g,t}(s) = \mathbb{1}\{t - e_g = s\}$ is an indicator for region g at time t being s periods away from its initial treatmnet, where $e_g = min\{t : w_{g,t} = 1\}$ denotes the first year of treatment for region g. Further, note that period $s = -1$ has been excluded, as it is common practice to normalize relative to the period immediately before the policy is introduced (Sun and Abraham, 2021).

Equation (1.2) can be estimated via Pooled Ordinary Least Squares (POLS). It can be shown that τ_s in (1.2) is an estimator of the ATT (see Wooldridge, 2021, for a discussion). In this context, the ATT is allowed to vary by treatment intensity, enabling the estimation of the long-term effects of RPs

The potential outcomes framework used to define the ATT in (1.1) implicitly encodes the **stable unit treatment value assumption (SUTVA)** (Roth et al., 2023).

For SUTVA to hold, two conditions must be met: i) *no interference*; ii) *treatment consistency*. The first assumes that the potential outcome of the region i is not affected by the treatment status of any other region $j \neq i$ (i.e., spillover effects are ruled out). The second requires that treatment effects be comparable across units. Specifically, if two units receive the same treatment, consistency implies that the policy should have the same impact on both units, all else being equal.

Additionally, in the program evaluation literature, when analyzing settings with multiple periods and variation in treatment timing, treatment is typically assumed to be at an **absorbing state**. That is, once the region g receives the treatment, it remains treated for the remainder of the panel (i.e., $w_{g,s} \leq w_{g,t}$ for $s < t$). The combination of SUTVA and the assumption that the treatment is at an absorbing state are sufficient to define the quantity in (1.1).

However, for τ_s in (1.2) to be a consistent estimator of the ATT, two additional assumptions must hold. The first rules out anticipatory behaviors.

No Anticipation (NA): For each treatment cohort $r \in \{2007, 2010, 2011\}$,

$$\mathbb{E}(y_t(r) - y_t(\infty)|d_{2007}, d_{2010}, d_{2011}) = 0, \quad \forall t < r \tag{1.3}$$

This means that, on average, the potential outcomes between treated and never-treated regions are the same in the pre-intervention period, regardless of when a region is first treated. This is similar to the strict exogeneity assumption required to estimate FE in panel data models.

The second assumption needed for identification is a generalization of the parallel trends assumption to the multi-period setup with staggered entry.

[20] The notation g is used instead of i for two reasons. First, it avoids confusion regarding the level of aggregation, as all variables are computed at the RHS level. Second, since regions represent the level at which standard errors will be clustered in the following analysis, this notation will align with that used in the cluster-robust literature (for instance, see Cameron and Miller, 2015).

Parallel Trends (PT): For each d_r with $r \in \{2007, 2010, 2011\}$

$$\mathbb{E}(y_t(\infty) - y_{first}(\infty)|d_{2007}, d_{2010}, d_{2011}) = \mathbb{E}(y_t(\infty) - y_{first}(\infty)) \tag{1.4}$$

where $t = \{y_{first}+1, \ldots, 2018\}$. This assumption requires the average evolution in the benchmark state to be mean independent of the treatment status. This must be true for every period relative to y_{first}.

Since it is unlikely that the PT and NA assumptions hold unconditionally (for a discussion, see Heckman et al., 1998; Abadie, 2005), the next regression will also be estimated:

$$\begin{aligned} y_{g,t} =& \alpha + \sum_r \lambda_r d_{g,r} + \sum_{s=-6}^{-2} \tau_s intens_{g,t}(s) + \sum_{s=0}^{11} \tau_s (w_{g,t} \cdot intens_{g,t}(s)) + \\ & \mathbf{x}_g \boldsymbol{\kappa} + \sum_r (d_{g,r} \cdot \mathbf{x}_g)\boldsymbol{\gamma}_r + (w_{g,t} \cdot \dot{\mathbf{x}}_{g,treat})\boldsymbol{\rho}_{treat} + \eta_t + u_{g,t} \end{aligned} \tag{1.5}$$

where $\mathbf{x}_g$ is a vector including a set of (time-invariant) covariates, $\dot{\mathbf{x}}_{g,treat} = (\mathbf{x}_g - \boldsymbol{\mu}_{treat})$, where $\boldsymbol{\mu}_{treat} = \mathbb{E}(\mathbf{x}_g|w_{g,t} = 1)$. The idea to center $\mathbf{x}$ about the mean of the x's over the treatment status ensures that τ_s represents the ATT (Wooldridge, 2021).[21]

The vector $\mathbf{x}_g$ includes all the socio-economic characteristics discussed in Section 1.5. However, since the TWM allows the inclusion of time-invariant covariates, the median values of the x's observed in the pre-treatment period for each region (i.e., before 2007) are used.

Unlike (1.2), in (1.5), the ATT is now allowed to vary not only across treatment intensities but also with covariates. However, in this case, standard errors should be adjusted to account for the sampling variation in $\boldsymbol{\mu}_{treat}$.[22] Allowing the ATT to vary with covariates enables researchers to relax both the PT and NA assumptions, requiring them to hold only within subpopulations with the same characteristics.[23] That is:

Conditional No Anticipation (CNA): For each treatment cohort $r \in \{2007, 2010, 2011\}$,

$$\mathbb{E}(y_t(r) - y_t(\infty)|d_{2007}, d_{2010}, d_{2011}, \mathbf{x}) = 0, \quad \forall t < r \tag{1.6}$$

Conditional parallel trends (CPT): For each d_r with $r \in \{2007, 2010, 2011\}$ and covariates $\mathbf{x}$,

$$\mathbb{E}(y_t(\infty) - y_{first}(\infty)|d_{2007}, d_{2010}, d_{2011}, \mathbf{x}) = \mathbb{E}(y_t(\infty) - y_{first}(\infty)|\mathbf{x}) \tag{1.7}$$

Regarding standard errors, it is common practice in the DiD literature to cluster them at the level at which the treatment is assigned (if known) to account for within-cluster correlation of the error term (Bertrand et al., 2004; Cameron and Miller, 2015; MacKinnon et al., 2023). As noted by Cameron and Miller (2015), including FE at the treatment level is insufficient to purge for all the intra-cluster correlation of the disturbances. Since the treatment assignment mechanism is known in this study, standard errors in regressions (1.2) and (1.5) will be clustered at the RHS level to correct for the potential serial correlation of the error term.

[21] Note it is sufficient to de-mean the x's only when interacted with $w_{g,t}$.

[22] This adjustment is implemented using the `vce(unconditional)` option in `margins`, which estimates standard errors while accounting for the sampling variation in the covariates.

[23] The TWM approach is similar to the estimator proposed by Sun and Abraham (2021). However, it offers a key advantage: the treatment effect is explicitly allowed to vary with covariates, making the parallel trends assumption more credible.

However, as explained in Section 1.3, the number of treated RHSs varies by year, with a maximum of 10. As shown by Cameron et al. (2008) and MacKinnon and Webb (2018), inference based on the classic Cluster-Robust Variance Estimator (CRVE) is highly likely unreliable when the number of treated clusters is small. To address this issue, both classic cluster-robust standard errors and those based on the *subcluster wild bootstrap* with Rademacher weights (MacKinnon and Webb, 2018) will be computed for each regression.

In this latter case, standard errors will be obtained using a version of the CRVE, and the resampling procedure will be performed at the region-year level. Each pseudo-residual contains only one point, the t^{th} observation of RHS g. The number of replications will be equal to $9,999$, as suggested by MacKinnon and Webb (2018).

Using subcluster wild bootstrap may improve finite-sample inference in settings where the number of treated clusters is small. Please refer to Appendix D for a review of how this procedure works.

1.6.2 Threats to Identification

For the estimator of the ATT obtained through TWM to be consistent, in addition to requiring treatment irreversibility and the validity of SUTVA, the No Anticipation and Parallel Trends assumptions must hold.

Specifically, TWM relies on a generalization of the PT and NA assumptions– commonly referred to as *invariance assumptions* in the literature of causal inference (Appendix B)– to construct $y_{g,t}(\infty)$. If either the PT or NA assumption fails, the estimator for the counterfactual outcome performs poorly, leading to inconsistency in the treatment effect estimator and unreliable inference.

Although invariance assumptions allow point identification of the treatment effect, they are difficult to justify in most empirical settings, as they require constructing the counterfactual with certainty.

In this context, while the assumption that treatment is at an absorbing state is a strong requirement– given that three regions (Liguria, Piemonte, and Sardegna) left treatment status–the NA assumption is likely more problematic. On the one hand, it is reasonable to assume that regions first treated in 2007 could not have anticipated the introduction of Recovery Plans. As explained in Section 1.2, previous measures aimed at curbing excessive local government spending had never been effectively implemented, making it unlikely that regions foresaw PdRs. On the other hand, justifying the absence of anticipatory effects for later-treated regions is more challenging.

For example, Calabria had already requested PdR activation in 2007. Still, the plan only became effective in December 2009 due to significant discrepancies between the region's balance sheets and those submitted to the Ministry of Health. The fact that Calabria had already required the activation of a plan in 2007 could indicate a potential violation of the NA assumption for Calabria.

To assess whether the NA assumption is likely violated and thus threatens the credibility of the TWM estimates, various sensitivity analyses addressing anticipatory behavior will be conducted (see Subsection 1.6.4).

The Parallel Trends assumption, on the other hand, is one of the most debated assumptions in applied work. In many settings, pre-existing trends are likely to be present before policy implementation. Several procedures have been proposed to test whether this assumption holds empirically. For example, Wooldridge suggests a method within the POLS framework to test the presence of pre-trends. This test, based on exclusion restrictions, introduces intensity dummies in the pre-treatment period ($s < 0$),

as I do in equations (1.2) and (1.5). Failing to reject the null hypothesis of joint insignificance for the coefficients of these intensity dummies will indicate the absence of pre-trends.

However, tests like the one proposed by Wooldridge (2021) are pre-tests for identifying assumptions. A substantial literature highlights several limitations of these tests. First, even if the PT assumption holds in the pre-treatment period, this does not guarantee it will hold post-treatment (Rambachan and Roth, 2023). Second, failing to reject the null hypothesis of absence of pre-trends could simply result from the low power of the test, rather than from the actual absence of such trends. This undermines inference, as the consistency of any DiD-type estimator relies on a valid PT assumption. Lastly, conditioning analysis on the test results–whether or not pre-trends are rejected–can induce a selection bias, known as *pre-test bias* (Roth, 2022).

Beyond the limitations of these tests, there is no consensus in the literature on how to proceed if pre-trends are detected. Therefore, the analysis should not be conditioned solely on the test results but also consider the researcher's context-specific knowledge regarding the likelihood of the PT assumption holding.

To assess whether the PT assumption is likely satisfied in this context, besides performing the test just mentioned above proposed in Wooldridge (2021), I plot, in Figure 1.2, the conditional average trajectory of each of the three treated cohorts against the never-treated group for each dependent variable of interest.

Before estimating conditional trends, I first applied a within-group transformation to the outcome of interest (Ferman and Pinto, 2016). The within transformation involved subtracting the mean of each dependent variable, computed over the period before the first PdR was signed (i.e., 2007). This ensures that any anticipatory effects from later-treated regions are excluded. The demeaned value is computed as follows:

$$y_{g,t} - \bar{y}_g \quad \text{where } \bar{y}_g = \frac{1}{T} \sum_{t=y_{first}}^{2006} y_{g,t} \text{ and } T = (2006 - y_{first} + 1) \tag{1.8}$$

Next, I regressed the outcome of interest on the socio-economic characteristics described earlier, along with year fixed effects, to obtain the conditional trends.

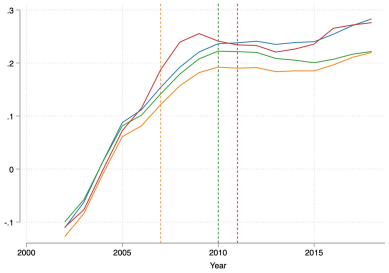
Figure 1.2 suggests that the PT assumption is likely satisfied for only a subset of the dependent variables, such as the mortality rate for ischemic heart disease or the inefficiency ratio.

The fact that there could be pre-existing trends, even after applying a within-transformation to the indicator of interest and conditioning on socio-economic characteristics, may be due to these controls being insufficient to "credibly" relax the PT and NA assumptions.

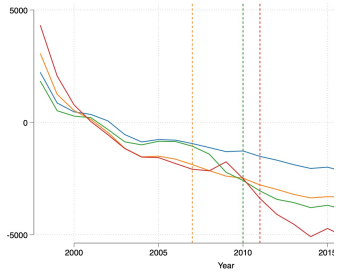
One way to address the potential invalidity of either the PT or NA assumptions is to rely on milder non-parametric assumptions, as proposed by Manski and Pepper (2013, 2018). Their estimator is robust to anticipatory behaviors and the presence of pre-trends, as will be explained in the following subsection. Specifically, to enhance the credibility of inference, Manski and Pepper (2013, 2018) suggest replacing invariance assumptions with *bounded variation* assumptions. These assumptions constrain the absolute difference between the true benchmark and the one constructed under an invariance assumption to a threshold, often denoted as δ, representing the level of uncertainty the researcher is willing to tolerate. The larger δ, the greater the permitted violation of the invariance assumption.

One last potential threat to the identification strategy could arise from violating the SUTVA. While

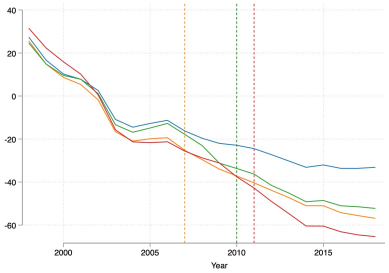
Figure 1.2: Time Series Plots – Evolution by Treatment Cohort



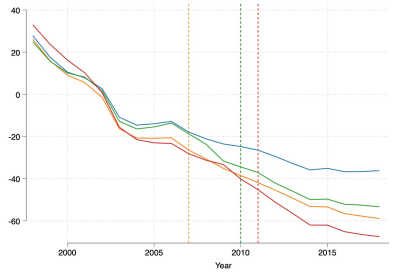
(a) ln(current health expenditure)



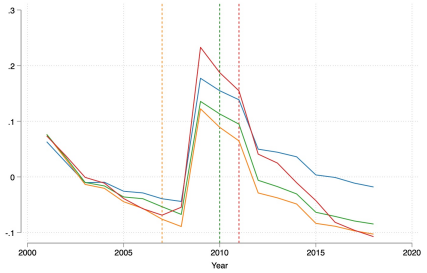
(b) Hospital beds



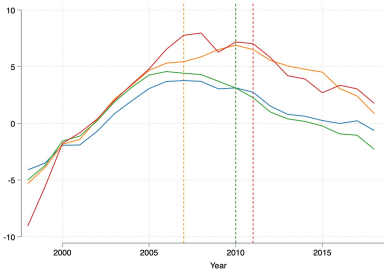
(c) Hospitalization rate



(d) Hospitalization rate (acute)

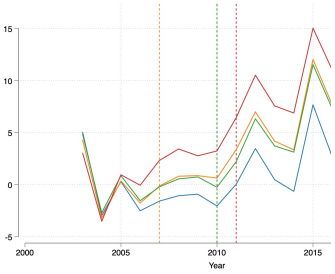


(e) Inefficiency ratio

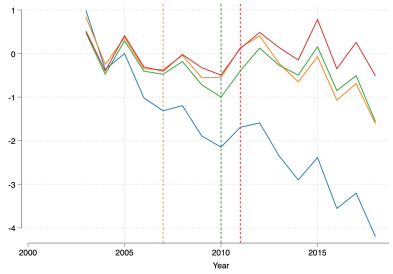


(f) % c-sections

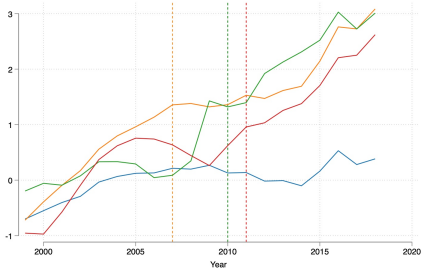
(g) % people with hip fracture replacement



(h) Mortality rate



(i) Mortality rate all causes



(j) % patient migrating for ordinary acute hospitalization

Notes: Results depicted in Panels (a) to (i) show the conditional average trajectory of each dependent variable of interest, split by treatment cohort. Specifically, the x-axis represents the years, while the y-axis shows the values of the indicator considered. The *blue* line represents the average trajectory for the never-treated regions (i.e., regions that never underwent a PdR during the observed period). The *orange* line shows the average trajectory for regions that first signed an RP in 2007 (Lazio, Abruzzo, Campania, Liguria, Molise, Sardegna, and Sicilia). The mean trajectory for regions entering an RP in 2010 (Calabria and Piemonte) is shown in *green*. The *red* line represents the trajectory of those RHSs that first signed an RP in 2011 (Puglia). The *orange dashed* vertical line marks 2007, the *green* dashed line marks 2010, and the *red dashed* vertical line marks 2011–the three years in which regions entered a PdR.

the assumption of no interference is likely valid, given that regions operate relatively independently in terms of healthcare service provision, treatment consistency might be violated. Specifically, one implicit assumption is that I do not distinguish between regions under external commissioners and those without.

One way to address this issue is to consider the presence of external commissioners as leading to different treatments and apply DiD methods that allow for multiple treatments, such as the one proposed by De Chaisemartin and d'Haultfoeuille (2023). However, due to the limited number of (treated) units in this analysis, allowing for further heterogeneity may not be feasible.

Nonetheless, the main goal of this paper is to evaluate the impact of Recovery Plans, regardless of whether region g at time t was under a commissioner. Thus, it should still be possible to define the quantity in (1.1). Additionally, this issue can be easily addressed by resorting to bounds. The bounds estimator allows the researcher to estimate region-specific treatment effects, distinguishing between regions under a commissioner and those not in a given year.

Furthermore, estimating region-specific treatment effects helps overcome the issue encountered when estimating (1.2) and (1.5), where I had to assume that the treatment was at an absorbing state even for the three regions that left the treatment status. Although "the effect of having ever received the treatment is of interest, as it captures the path of treatment effects even though the treatment itself may be transient" (Sun and Abraham, 2021, p. 177), assuming that the region behaved as if it were still under treatment may be too strong. If one considers this assumption too stringent, the estimated bounds for the ATT for these three regions can be disregarded, as explained further below.

1.6.3 Bounds

To provide intuition behind the non-parametric approach proposed by Manski and Pepper (2018), I follow Depalo (2019). Suppose the counterfactual outcome, $y_{g,t}(\infty)$, must be estimated. In the program evaluation literature, this is typically retrieved by invoking an "exact" invariance assumption (Imbens and Wooldridge, 2009; Depalo, 2019). As noted in Depalo (2019), these assumptions can be categorized into four main groups.

The first is the *time invariance* assumption, which exploits the outcome observed in the pre-treatment period for the treated region g to estimate the benchmark outcome:$\hat{y}_{g,t}(\infty) = y_{g,r-1}(\infty)$ where $y_{g,r-1}$ denotes the value of y for region g in period $r-1$ (the last available pre-treatment period).

Another common approach is the *state invariance* assumption, which utilizes (often a linear combination of) observed outcomes in never-treated regions. If $never$ denotes the set of never-treated regions, then for $t \geq r$, one way to retrieve $\hat{y}_{g,t}(\infty)$ is: $\hat{y}_{g,t}(\infty) = y_{never,t}(\infty)$ or $\hat{y}_{g,t}(\infty) = \mathbb{E}(y_{never,t}(\infty))$.

Alternatively, a *parallel trends* assumption can be employed, as discussed in Section 1.6.1. Lastly, the *time-varying parallel trends* assumption retrieves the counterfactual, $y_{g,t}(\infty)$, using a weighted average of regions in the donor pool, with weights chosen according to the Synthetic Control (SC) Method (Abadie and Gardeazabal, 2003): $\hat{y}_{g,t}(\infty) = \mathbf{w} y_{never,t \geq r}(\infty)$ where $\mathbf{w}$ is the vector of selected weights.[24]

If any invariance assumptions hold, the benchmark outcome can be identified, thus the treatment effect. However, if an exact invariance assumption does not hold, Manski and Pepper (2013, 2018)

[24]For a concise overview of deriving the counterfactual outcome and estimating the average treatment effect, refer to Appendix B.

suggest explicitly accounting for uncertainty by allowing the invariance assumption to hold only approximately. This means that the benchmark against which $y_{g,t}(r)$ should be compared will be at most equal to the counterfactual estimated based on the suitable invariance assumption plus the degree of uncertainty considered: $\hat{y}^{Up}_{g,t}(\infty) = \hat{y}_{g,t}(\infty) + \delta$. On the other hand, the counterfactual outcome will be at least equal to the estimated counterfactual minus the level of uncertainty allowed: $\hat{y}^{Low}_{g,t}(\infty) = \hat{y}_{g,t}(\infty) - \delta$.

This approach identifies bounds for $y_{g,t}(\infty)$ (Manski, 1990). Although δ need not be symmetric in principle, assuming $\delta^{Up} \neq \delta^{Low}$ "would constrain the spectrum of possible answers" (Depalo, 2019, p. 6) in terms of economic models one wishes to test. Thus, the analysis will consider only symmetric δ.

Several factors can lead to violations of exact invariance assumptions, such as pre-trends between treated and control regions, omitted variables, or anticipatory behavior. However, a key advantage of the estimator proposed by Manski and Pepper (2018) is that the specific reason for violation does not need to be known for the ATT estimator to be consistent–only that the level of uncertainty, δ, must be specified ex-ante. The larger δ, the weaker the reliance on the invariance assumption and the greater the credibility of the estimator (Manski, 2003).

Since δ is unobservable, one approach to selecting its value is to set it equal to the absolute difference observed in the pre-treatment period between the actual outcome at period t, $y_{g,t}$, and the benchmark constructed according to one of the exact invariance assumptions presented above, $\hat{y}g, t(\infty)$. Manski and Pepper (2018), for instance, use the 75th percentile of this difference or the maximum observed deviation between yg, t and $\hat{y}_{g,t}(\infty)$. Following Depalo (2019), this analysis sets δ equal to the largest pre-treatment difference observed before a PdR was introduced. The reason to favor a larger value of δ is that increasing the level of uncertainty strengthens the credibility of the findings:

$$\delta^{obs} \equiv \max_{t \in \{y_{first}, \ldots, r-1\}} \{|y_{g,t} - \hat{y}_{g,t}(\infty)|\} \quad g = 1, \ldots, 21 \tag{1.9}$$

Where $\hat{y}_{g,t}(\infty)$ is estimated using one of the four exact invariance assumptions discussed earlier.

To illustrate this approach, consider the following example. Suppose we want to estimate the counterfactual outcome for region i and have four pre-treatment periods, $t = 1, \ldots, 4$. Assume the observed outcomes are $y_1 = 10$, $y_2 = 11$, $y_3 = 9$, and $y_4 = 12.3$. Suppose further the counterfactual outcomes are estimated via SC: $\hat{y}_{SC,1} = 8$, $\hat{y}_{SC,2} = 11.5$, $\hat{y}_{SC,3} = 8$, and $\hat{y}_{SC,4} = 9$. Then, according to (1.9), $\delta = 2.3$, which corresponds to the largest absolute difference between the observed outcome y_t and the counterfactual estimate $\hat{y}_{SC,t}$.

In particular, increasing δ enhances credibility and widens the bounds, reducing precision. I acknowledge this credibility-precision trade-off; however, I favor larger bounds over more precise estimates for this analysis to ensure robustness. Specifically, setting $\hat{\delta}$ as in (1.9) ensures that the estimated bounds for the treatment effect–and, consequently conclusions drawn from them–remain robust even when accounting for the largest observed violation of the exact invariance assumption in the pre-treatment period.

By varying δ, a full range of possible counterfactual outcomes is identified. The treatment effect for region g undergoing an RP in period r, computed at time t, is then bounded by:

$$\tau^{Low}_{g,r,t} = \mathbb{E}(y_{g,t}(r) - \hat{y}^{Up}_{g,t}(\infty)|d_r = 1)$$

$$\tau^{Up}_{g,r,t} = \mathbb{E}(y_{g,t}(r) - \hat{y}^{Low}_{g,t}(\infty)|d_r = 1)$$

for $r \in 2007, 2010, 2011$ and $t = r, \ldots, 2018$. If both $\tau^{Low}_{g,r,t}$ and $\tau^{Up}_{g,r,t}$ are positive, then the region-specific treatment effect, $TE_{g,r,t}$, is positive; if $\tau^{Low}_{g,r,t}$ and $\tau^{Up}_{g,r,t}$ are both negative, then the $TE_{g,r,t} < 0$; if $[\tau^{Low}_{g,r,t}, \tau^{Up}_{g,r,t}]$ covers 0, it is not possible to say anything about the $TE_{g,r,t}$.[25]

Another benefit of exploiting assumptions of bounded variation is that they allow combining more assumptions to obtain a refinement of the length of $[\tau^{Low}_{g,r,t}, \tau^{Up}_{g,r,t}]$. This also avoids favoring one particular assumption. Following Depalo (2019), time invariance and Synthetic Control will be jointly used to estimate the causal effect of Piani di Rientro. The SC assumption is preferred over state-invariance or DiD assumptions because it constructs the benchmark using all available pre-treatment information rather than a simple average of never-treated regions (state-invariance) or only the last pre-treatment period (PT). Further, how SC constructs the counterfactual is similar to how the TWM does. However, unlike Depalo (2019), a longer time horizon is considered to estimate long-term PdR impacts.

If $\delta_{time} \neq 0$ and $\delta_{SC} \neq 0$ denote the uncertainty parameters of the time invariance and SC assumptions, respectively, then the bounds for $y_{g,t}(\infty)$ can be obtained as follows:

$$y^{Low}_{g,t}(\infty) \equiv max(y_{SC,t}(\infty) - \delta_{SC}, y_{r-1,t}(\infty) - \delta_{time}) \leq y_{g,t}(\infty) \leq$$
$$y^{Up}_{g,t}(\infty) \equiv min(y_{SC,t}(\infty) + \delta_{SC}, y_{r-1,t}(\infty) + \delta_{time})$$

For $y^{Low}_{g,t}(\infty) \leq y_{g,t}(\infty) \leq y^{Up}_{g,t}(\infty)$ to hold, a necessary condition would be that $\delta_{SC} + \delta_{time} \geq |y_{SC,t}(\infty) - y_{r-1,t}(\infty)|$. However, such a condition is not imposed in the following analysis, as being unable to identify the bounds for the treatment effect might also be relevant to draw policy implications. The intuition is that if the upper bound is lower than the lower bound, nothing can be inferred about the treatment effect in such a situation.

Regarding statistical inference, this study does not explicitly address how to perform inference in this context. The primary reason is that the data used represent the entire population rather than a random sample. Moreover, Manski and Pepper (2018, p. 234) argue that "a fundamental reason for not performing statistical inference is that measurement of statistical precision requires specification of a sampling process that generates the data. Yet we are unsure what type of sampling process would be reasonable to assume in this application".

One way to address this issue would be to view current Italy as the realization of a sampling process. However, this would require defining a super-population and a stochastic process governing Italy's actual history–an approach for which no consensus exists in the current literature on bounds.

For instance, Rambachan and Roth (2023) propose two methods for obtaining uniformly valid inference by placing restrictions on potential post-treatment violations of the parallel trends assumption. Their approach assumes that the estimator used to construct the counterfactual is asymptotically normally distributed. However, as discussed in Appendix D, with only a few treated clusters, asymptotics do not hold, and the Central Limit Theorem is inapplicable. This is why the present analysis favors the estimator proposed by Manski and Pepper (2013, 2018).

[25] With bounded variation assumptions, treatment effects for each treated region can be estimated, offering a significant advantage. Moreover, note that by exploiting the linearity of $\mathbb{E}(\cdot)$ and the fact that we are now considering the treatment effect for each treated region in each instant of time, $\tau^{j}_{g,r,t} = \mathbb{E}(y_{g,t}(r) - \hat{y}^{j}_{g,t}(\infty)|d_r = 1) \equiv y_{g,t}(r) - \mathbb{E}(\hat{y}^{j}_{g,t}(\infty)|d_r = 1)$ with $j \in \{Low, Up\}$.

Another advantage of using bounds instead of the approach by Rambachan and Roth (2023) is that the former provides a robust alternative to the TWM, as it remains valid even when both the parallel trends (PT) and no anticipation (NA) assumptions are violated.[26]

Bei (2024) alternatively introduces confidence intervals for union bounds using a modified conditional inference approach when the union is taken over a finite set. However, this approach also requires the estimator used to construct the counterfactual outcome to be asymptotically Normally distributed.

1.6.4 Robustness Checks

Regressions (1.2) and (1.5) represent just two possible specifications of the TWM approach. Greater heterogeneity could be introduced by employing the estimator proposed by Wooldridge. For example, the ATT could vary over time and across treatment cohorts. In addition, the treatment dummy can also be interacted with controls, cohorts, and year dummies (see, for instance, Eq. (B.11) in Appendix B). This approach offers greater flexibility, allowing for a more relaxed version of the PT and NA assumptions compared to (1.5).

However, in this context, where the number of observations is fixed and small while T is large, rather than allowing the treatment effects to vary across treated cohorts and over time and then aggregating them ex-post, the ATT was allowed to vary only over time and covariates.

In the following analysis, another version of the TWM will also be estimated as a robustness check. Specifically, the ATT will be allowed to vary by calendar time, while homogeneity across cohorts will be imposed.[27]

The calendar-type versions of the TWM that will be estimated are the following:

$$y_{g,t} = \alpha + \sum_{s=2007}^{2018} \tau_s(w_{g,t} \cdot fs_t) + \sum_r \lambda_r d_{g,r} + \eta_t + u_{g,t} \tag{1.10}$$

$$\begin{aligned} y_{g,t} =& \alpha + \sum_{s=2007}^{2018} \tau_s(w_{g,t} \cdot fs_t) + \sum_r \lambda_r d_{g,r} + \sum_r (d_{g,r} \cdot \mathbf{x}_g)\boldsymbol{\gamma}_r + \mathbf{x}_g \boldsymbol{\kappa} \\ &+ (w_{g,t} \cdot \dot{\mathbf{x}}_{g,treat})\boldsymbol{\rho}_{treat} + \eta_t + u_{g,t} \end{aligned} \tag{1.11}$$

The only difference between this version of the TWM and (1.2) and (1.5) lies in the fact that $w_{g,t}$ is now interacted with fs_t (instead of $intens_{g,t}(s)$), where fs_t is a dummy variable equal to 1 if $s = t$ and zero otherwise. Compared to (1.2) and (1.5), $fs_t = 1$ tells how much being treated in a specific year (e.g., 2008) affects the dependent variable of interest. This version of the TWM is closer to bounds, as the ATT will be allowed to vary by calendar time, rather than by how many periods have passed since the region first received the treatment.

The TWM is just one of several approaches recently proposed to estimate the ATT in a context with variation in treatment timing (see Appendix B). To assess the robustness of the results obtained through the TWM, the estimator proposed by Sun and Abraham (2021) will also be used. This method

[26] As a robustness check, the method by Rambachan and Roth (2023) is also applied. However, under the assumption that $M = 0$–meaning no post-treatment violations of PT–the confidence intervals derived from this approach are much wider than those obtained using TWM or the estimator from Sun and Abraham (2021). This further supports the idea that normality is unlikely to hold in this scenario.

[27] As an additional sensitivity exercise, the ATT was also allowed to vary by cohorts, but no clear pattern was detected across cohorts. This result may be driven by the fact that there are few treated regions in the second and third cohorts, causing the standard errors to be imprecisely estimated.

provides an estimator for τ_s, similar to the one obtained by estimating Eq. (1.2) and (1.5) via POLS. In doing so, Sun and Abraham (2021) exploit a generalization of the PT and NA assumptions similar to those presented in (1.3) (1.4). However, compared to the estimator proposed by Wooldridge (2021), this method has one significant drawback: the ATT is not explicitly allowed to vary with covariates, making the parallel trend and no-anticipation assumptions less likely to be valid.

Regarding bounds, as anticipated in Subsection 1.6.3, the largest values of δ observed in the pre-treatment periods will be used. This approach is consistent with the idea that the greater the uncertainty allowed, the more credible the results (Manski, 2003). However, to assess the robustness of the findings, bounds will also be re-estimated by setting δ equal to the 75^{th} percentile of the (absolute) difference observed in the pre-treatment period between the actual outcome and the benchmark. As a further sensitivity exercise, a complete set of combinations of δ_{SC} and δ_{time} will be tested. In this latter case, whether the time invariance and SC assumptions jointly hold in the data can be directly tested. To give the intuition, suppose that δ_{SC} and δ_{time} are set jointly equal to 0. If the upper bound is always smaller than the lower bound –meaning that the ATT is not identified– this would imply that we can reject the hypothesis that the time invariance and SC assumptions jointly hold in the data.

As discussed in Subsection 1.6.2, while it is reasonable that regions that first received the treatment in 2007 could not have anticipated the introduction of PdRs, it is less plausible that later-treated regions did not foresee their eventual entrance in a PdR. Although bounds allow retrieving a consistent estimator of the ATT without knowing why the invariance assumption fails, various sensitivity exercises will be conducted to assess whether the (C)NA assumption is likely to hold.

First, following Wooldridge (2021), leads of the treatment dummy will be introduced in Eq. (1.10) and (1.11). If the null hypothesis of joint insignificance of the coefficients for the treatment dummy leads cannot be rejected, this would suggest that the (C)NA assumption is likely to hold. However, as with the PT assumption, this may introduce pre-test bias (Roth et al., 2023). To mitigate this issue, and in line with the approach proposed by Callaway and Sant'Anna (2021), equations (1.10) and (1.11) will be re-estimated by artificially anticipating the entry of each region into the PdR by two years (e.g., treating Abruzzo as if it had joined an RP in 2005). This two-year anticipation is motivated by the fact that an initial draft of Recovery Plans was outlined in the budget law for 2005, although this measure has never become effective.

Furthermore, following Leive et al. (2023) and McKibbin (2023), who exclude later-treated units for which they expect the NA to be potentially violated from their analysis, all analysis in Subsection 1.6.1 will be re-run while excluding later-treated regions (Calabria, Piemonte, and Puglia). Precisely, the following two regressions will be estimated:

$$y_{g,t} = \alpha + \sum_{s=0}^{11} \tau_s (w_{g,t} \cdot intens_{g,t}(s)) + \lambda d_{g,2007} + \eta_t + u_{g,t} \tag{1.12}$$

$$\begin{aligned} y_{g,t} =& \alpha + \lambda d_{g,2007} + \sum_{s=0}^{11} \tau_s (w_{g,t} \cdot intens_{g,t}(s)) + \mathbf{x}_g \boldsymbol{\kappa} + (d_{g,2007} \cdot \mathbf{x}_g)\boldsymbol{\gamma} + \\ & (w_{g,t} \cdot \dot{\mathbf{x}}_{g,treat})\boldsymbol{\rho}_{treat} + \eta_t + u_{g,t} \end{aligned} \tag{1.13}$$

Additionally, calendar-type versions of the TWM–such as those specified in (1.10) and (1.11)–will also be estimated. To save space, these results are available upon request.

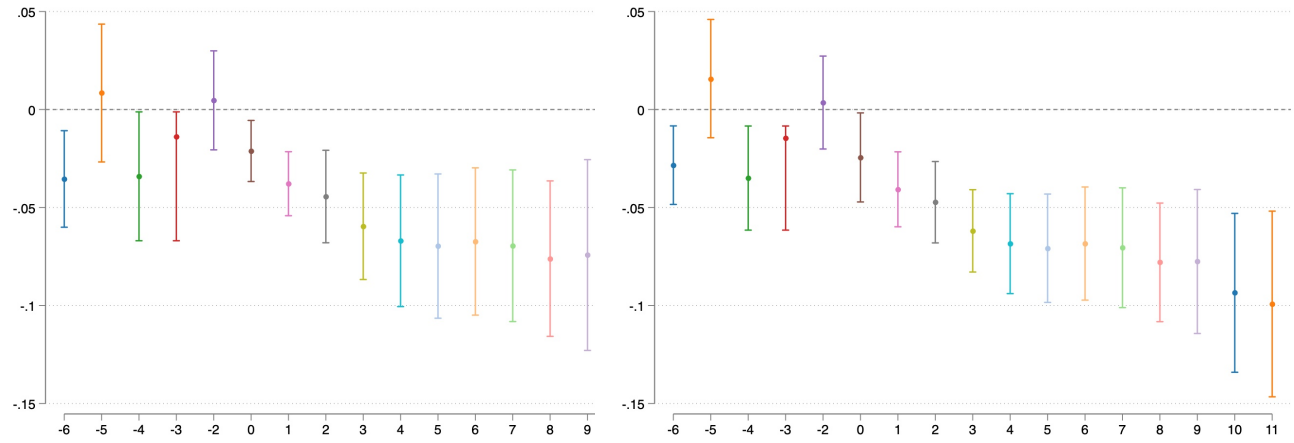
Lastly, to assess the validity of treatment irreversibility assumption, I will re-estimate both equations (1.2) and (1.5), excluding regions that exited the treatment status during the observed period: Liguria, Sardegna, and Piemonte.[28]

1.7 Results

1.7.1 Results – Two-Way Mundlak

This subsection presents the results obtained using the Two-Way Mundlak approach. The decision to use a method that explicitly accounts for the staggered nature of the policy is motivated by the DiD decomposition. Since the results of the DiD decomposition align with theoretical predictions, they are omitted from the main text and reported only in Appendix C.

Figure 1.3: Results Two-Way Mundlak – ln(Current health exp)



(a) Without controls

(b) With controls

Notes: The above regressions include 357 RHS-year observations. Results depicted in Panel (a) were obtained by estimating (1.2) using as the dependent variable the log of current health expenditure. In Panel (b) are reported the results obtained by estimating (1.5). Coefficients are reported with 95% confidence intervals obtained via the subcluster wild bootstrap (MacKinnon and Webb, 2018) with Rademacher weights and $9,999$ replications. Specifically, the t-statistic is obtained through a CRVE estimator (where the level of clustering is at the RHS level), whereas the resampling is carried out at the RHS-year level.

Panels (a) and (b) in Figure 1.3 show the estimated coefficients for $w_{g,t} \cdot intens(s)$, obtained by estimating equations (1.2) and (1.5), respectively, with the log of current health expenditure as the dependent variable. These are standard event-study plots, where the y-axis represents the estimated coefficients and the x-axis shows the number of periods that have passed since region g first received the treatment.[29]

[28] To address the issue that some units experience treatment switching on and off at different points in time, I also employ the event-study estimator proposed by De Chaisemartin and d'Haultfoeuille (2022a), which allows for non-absorbing treatment. However, results for this estimator are omitted to save space and because it requires normality of the error term–an assumption that is unlikely to hold in this context, as previously discussed.

[29] For example, $w_{g,t} \cdot intens(0)$ represents the period immediately after the region first received the treatment (for Abruzzo, treated in 2007, $w_{g,t} \cdot intens(0)$ is equal to 1 in 2007, while for Calabria, which was treated in 2010, it is equal to 1 in 2010). Conversely, $w_{g,t} \cdot intens(5)$ represents the case where 5 years have passed since treatment was first applied (for Abruzzo, this would correspond to 2012).

As discussed earlier, the coefficients are reported with 95% confidence intervals, obtained using the subcluster wild bootstrap with Rademacher weights (MacKinnon and Webb, 2018).[30]

The results presented in Figure 1.3 are complemented by those in Table E.1 in Appendix E, which aggregates the coefficients from the figure. Specifically, I present three average effects: short-term, long-term, and overall ATT. *Short-term* refers to the average effect over the first three years of the PdR (i.e., the average of τ_s for $s \in 0, 1, 2$). *Long-term* represents the average effect over the last three observed years, when the RHSs are 10 to 12 years after the PdR was first signed. *Overall* represents the average effect of the PdRs across all post-treatment periods. Grouping the results helps highlight temporal trends and estimate the average overall effect of PdRs. Additionally, Table E.1 reports the test for the presence of pre-trends, as proposed by Wooldridge (2021) and discussed earlier. To save space, only the results obtained by conditioning on covariates are presented in this table.

On average, the central government successfully reduced current health expenditures in treated regions with the introduction of RPs, as expected. Except for the coefficient for $intens(0)$, the effect is consistently negative and statistically significant at the 5% level across both panels. One possible explanation for failing to reject the null hypothesis that the coefficient for $intens(0)$ is equal to zero is that regions may need time to adjust to the PdR-specific requirements before they can begin reducing current health expenditures. However, the average short-term effect of the policy is negative and statistically different from zero.

While the stability of the results across the two specifications is reassuring, comparing Panel (a) with Panel (b) reveals that conditioning on covariates only marginally affects the results, with estimated coefficients being similar in magnitude across both panels. This suggests that the Parallel Trends and No Anticipation assumptions may not be fully relaxed by controlling for only a few covariates. Alternatively, this could reflect the absence of covariate-specific trends for this variable.

Furthermore, the statistical significance of some coefficients for $w_{g,t} \cdot intens(s)$ for $s < 0$ may point to the presence of pre-trends, a finding supported by the pre-trend test, which rejects the null hypothesis of joint insignificance for the coefficients of $w_{g,t} \cdot intens(s)$ for $s < 0$. However, as noted earlier, there are limitations to these types of tests (Rambachan and Roth, 2023). Still, as shown in Figure 1.2, it is likely that the PT assumption holds for this dependent variable.

Before presenting the results for the other indicators, it is important to emphasize that the subcluster wild bootstrap procedure performs well in this setting. Further details on this procedure can be found in Appendix D.

The results for the other indicators discussed in Section 1.5 are presented in Figure 1.4 and Table E.2. To save space, only the results obtained by conditioning on covariates are reported (i.e., by estimating (1.5)). The intuition behind these results is similar to that for the log of current health expenditure, as conditioning on covariates only marginally affects the ATTs.

Regarding the two main targets, the results confirm that, as expected, the introduction of Recovery Plans led to a reduction in hospitalization rates–both total and acute hospitalizations (Panels (a) and (b), respectively)–as well as a decrease in the number of hospital beds (Panel (c)) in treated regions. For each dependent variable, the estimated coefficients for $w_{g,t} \cdot intens(s)$ are always statistically different

[30] Similar results were obtained using standard errors calculated with a classic CRVE estimator. The estimated τ_s's are consistently statistically different from zero at the 5% level, except for the coefficient of $intens(0)$. However, the confidence bands are now slightly wider than before. This finding is consistent with Monte Carlo simulation results in MacKinnon and Webb (2018), which suggest that the classic CRVE may perform poorly in such contexts. These results are presented in Figure E.1 in Appendix E.

from zero, except during the first two years following implementation (and for $w_{g,t} \cdot intens(11)$ in the case of hospital beds).

The absence of a significant effect in the initial years is in line with the policy's nature, which requires a structural reorganization of the RHS. This suggests that regions need time to fully adapt to the Recovery Plan requirements before achieving significant reductions in hospitalization rates and the number of beds. While the short-term effect for these indicators is negative and statistically significant, its magnitude is relatively small compared to the long-term impact, reinforcing the idea that adjustments take time.

It is also worth noting that while Figure 1.4 shows that the estimated coefficients for the interaction term always include zero for all three indicators, we reject the null hypothesis of no pretrends for the number of hospital beds.[31] This finding is corroborated by the time-series evolution depicted in Panel (b) of Figure 1.2, where treated regions in 2007 and 2011 appear to follow a different trend compared to the never-treated regions.

Regarding inefficiency (Panel (d) in Figure 1.4), we fail to reject the null hypothesis of absence of pre-trends, consistent with Figure 1.2. Following the introduction of PdRs, treated regions experienced a small but statistically significant reduction in inefficiencies, as measured by the ratio between the hospitalization for DRG at high risk of inappropriateness versus those that could not be avoided. A lower ratio indicates lower inefficiency in the RHS.

For this outcome, the estimated coefficients for $w_{g,t} \cdot intens(s)$ are statistically significant at the 5% in most cases. However, when considering the c-section rate (Panel (e)), the coefficients for $w_{g,t} \cdot intens(s)$ are always indistinguishable from 0. Nonetheless, Table E.2 reveals a positive and statistically significant short-run and overall effect, while the average long-run effect is not distinguishable from zero–which is puzzling.

These findings suggest that PdRs may have contributed to reducing inefficiencies in the healthcare system.

While PdRs successfully reduced costs, they also appear to have deteriorated the quality of healthcare services. This decline is observed across different quality indicators, except the percentage of patients with hip fracture replacement (Panel (h)).

Before discussing the other quality indicators, it is essential to consider why no effect is detected for the hip fracture replacement rate. There are two potential explanations. From a contextual perspective, this indicator falls under the *specific targets* of PdRs. Since regions and hospitals have a financial incentive to ensure that patients hospitalized with a hip fracture undergo surgery within 48 hours of arriving at the emergency room, they may have prioritized maintaining performance on this metric. From an econometric perspective, the ATT estimators may be inconsistent. Although we fail to reject the null hypothesis of joint insignificance, Figure 1.2 suggests that the Parallel Trends assumption is likely to be violated for this indicator.

Regarding the other proxies of quality, the introduction of PdRs appears to have led to an increase in both the all-cause mortality rate (Panel (f)) and the mortality rate from ischemic heart diseases (Panel

[31] The discrepancy between the test of joint significance of the τ_s coefficients and the results depicted in Figure 1.4 is not unexpected. In Figure 1.4, we are testing the null hypothesis $H_0 : \tau_s = 0$ individually for each τ_s, whereas the joint significance test evaluates whether all the τ_s's are simultaneously equal to zero. As noted by Davidson et al. (2004), the t-statistics used for individual hypothesis testing rely solely on the marginal distribution of $\hat{\tau}_s$. At the same time the F-test for joint significance accounts for the joint distribution of the estimators of the τ_s's. Since the τ_s's are likely correlated, the t-statistics fail to account for this dependency, which may explain the difference in results between the two approaches.

(g)). Furthermore, for both indicators, the long-term effect is larger than the short-term effect. Specifically, for the mortality rate from ischemic heart diseases, the coefficients for $w_{g,t} \cdot intens(s)$ are statistically significant in almost all cases, with point estimates ranging from barely 1 to almost 3 p.p.

However, it is essential to note that while from Figure 1.2 the conditional Parallel Trend seems to hold for both indicators, the formal test rejects the null hypothesis for all-cause mortality.

The rise in all-cause mortality observed in treated regions after the implementation of PdRs aligns with previous research linking reductions in healthcare spending to higher mortality rates (e.g., Golinelli et al., 2017). Likewise, the increase in ischemic heart disease mortality may be attributed to factors such as decreased hospital funding (Kwok et al., 2018; Lobo et al., 2020), diminished hospital capacity (Brown et al., 2014; Boden et al., 2016), or longer travel times due to hospital closures (Ghislandi et al., 2025).

Regarding the impact of the policy on the percentage of patients migrating for ordinary acute hospitalizations (Panel (i)), while Figure 1.2 suggests that the conditional parallel trends assumption may not hold, the formal test fails to reject the null hypothesis of joint insignificance. A statistically significant increase is observed in regions that are at least four periods after the first time they were treated.

This documented increase in patient migration in treated regions aligns closely with the findings of Beraldo et al. (2023), as well as with existing studies showing that healthcare spending cuts (Arcà et al., 2020) and reductions in hospital capacity (Balia et al., 2018) tend to induce higher patient mobility. The fact that the coefficients for $w_{g,t} \cdot intens(0)$ to $w_{g,t} \cdot intens(3)$ are statistically indistinguishable from zero is consistent with the short-term effect not being statistically significant. This result may be due to the time required for regions to fully adjust to the PdR's specific targets, as well as the additional time it takes for patients to perceive changes in service quality.

Although the increase in mortality rates (both all-cause and ischemic heart disease-related) may represent a direct negative consequence of the policy, the observed increase in patient mobility is likely driven by a perceived decrease in the quality of healthcare services in the patient's region of origin.

1.7.2 Results – Bounds

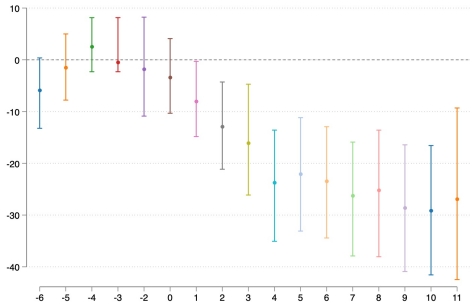
Given that the Conditional PT and NA assumptions may be violated for some indicators, we complement the TWM results with those obtained using a set of bounded variation assumptions.

The covariates used to construct the benchmark via the Synthetic Control are the same as those described in Section 1.5. However, before applying any invariance assumptions to compute bounds on $\tau_{g,r,t}$, each indicator is first within-transformed as in (1.8) (Ferman and Pinto, 2016).

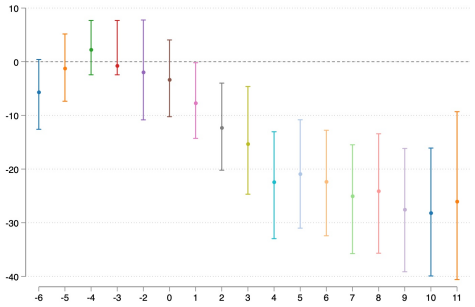
Table F.3 of Appendix F reports, for each treated region, the difference between the observed outcome and the estimated counterfactual built invoking either Synthetic Control or time invariance assumptions, up to the last year before the introduction of the RP. As explained earlier, the optimal value of δ is set equal to the largest absolute difference observed in the pre-treatment period between the actual outcome and the estimated benchmark. However, if the largest absolute difference for a given indicator is less than or equal to .1, δ is increased by 0.1 for all regions.

Table 1.3 present the bounds on the $\tau_{g,r,t}$ for costs indicators based on the optimal values of δ (i.e., $\delta_{T,Max}$, $\delta_{SC,Max}$). The corresponding results for efficiency indicators are shown in Table 1.4, while those for quality proxies are reported in Table 1.5. Specifically, results for each treated region are displayed in columns, while results for each post-treatment period appear in rows.

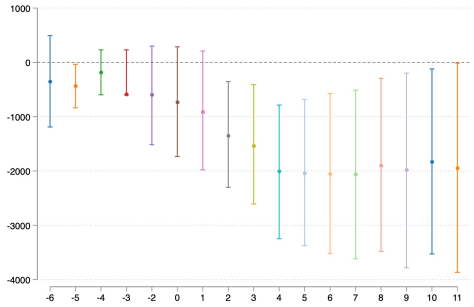
Figure 1.4: Results Two-Way Mundlak – Other indicators



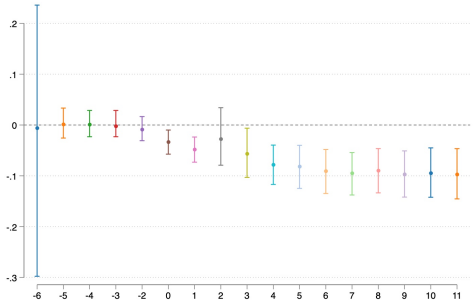
(a) Hospitalization rate



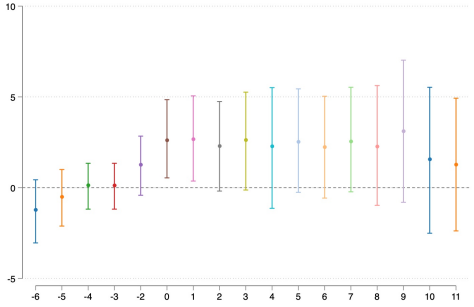
(b) Hospitalization rate – acute



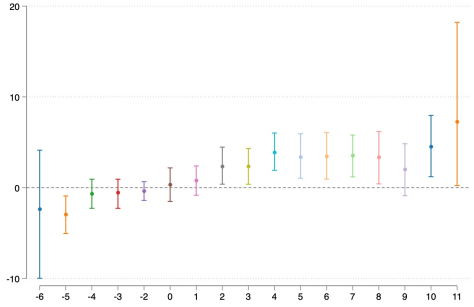
(c) Hospital beds (ordinary regime)



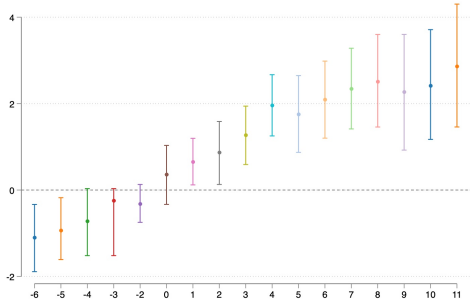
(d) Inefficiency ratio



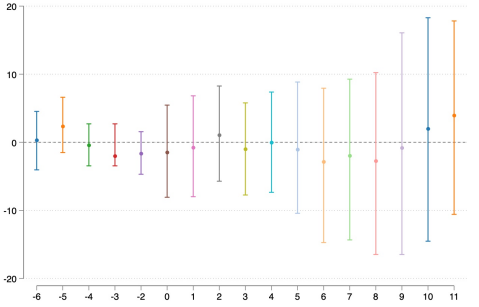
(e) % C-sections



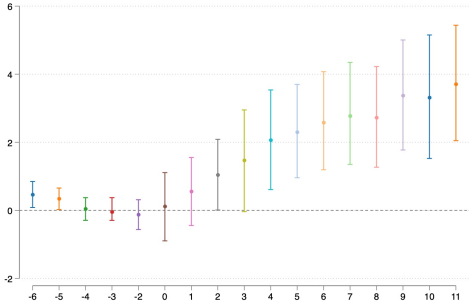
(f) Mortality rate



(g) Mortality rate from ischaemic heart diseases



(h) % patients with hip fracture replacement



(i) % patients migrating for ordinary acute hospit.

Notes: Results depicted in Panel (a) to (i) were obtained by estimating (1.5) using as the dependent variable the name of the corresponding Panel. Coefficients are reported with 95% confidence intervals obtained via the subcluster wild bootstrap (MacKinnon and Webb, 2018) with Rademacher weights and 9,999 replications. Specifically, the t-statistic is obtained through a CRVE estimator (where the level of clustering is at the RHS level), whereas the resampling is carried out at the RHS-year level.

Before discussing the findings, it is essential to clarify that missing values observed for Calabria and Piemonte (2007-2010) and Puglia (2007-2011) across all indicators stem from these regions entering treatment status in 2010 and 2011, respectively. Any other missing values occur when the estimated upper bound is lower than the lower bound. For instance, in the case of hospital beds, Liguria exhibits four missing values–one in 2014 and three in 2016-2018–because, in these years, the estimated upper bound for $\tau_{g,r,t}$ falls below the lower bound.

Regarding the results obtained using the log of current health expenditure as the dependent variable, surprisingly, there seems to be no effect. Except for Liguria (2016-2018) and Sicilia (2011), where both the upper and lower bounds are below zero–indicating a negative treatment effect–the bounds for all other regions include zero. In these cases, it is impossible to draw conclusions about the treatment effect. On the other hand, a positive effect is observed for Sardegna (2012, 2014-2018), which is counterintuitive, as one would expect a decrease in health expenditure following the introduction of a PdR. However, Sardegna is assumed to be at an absorbing state, even if it exited treatment status in 2010 due to its special statute. A similar argument applies to Liguria and Piemonte, which also left the treatment status after meeting the plan's goals.

As mentioned earlier, bounds have the advantage of allowing for the estimation of the treatment effect separately for each treated region. Therefore, if the assumption that these regions are at an absorbing state is too demanding is considered, the columns for Liguria, Piemonte, and Sardegna can be ignored.

Overall, the absence of a clear effect on the log of current health expenditure can be explained by two reasons. First, as shown in Figure 1.2, the introduction of RPs did not lead to a decrease in healthcare spending for treated cohorts. Instead, it resulted in a flattening of the trend. Furthermore, the trend experienced by treated regions closely parallels that of the never-treated RHS. Second, Figure E.2–presented in the following subsection–reports the estimated coefficients of τ_s obtained by estimating equations (1.11), using the log of current health expenditure as the dependent variable (this version of the TWM is closer to the bounds approach). Although a negative and statistically significant effect is detected, the point estimates are very small in magnitude. This suggests that the absence of a significant effect may be due to the ATT, which was already small, being further split into ten different region-specific treatment effects.

Regarding the impact of Recovery Plans on hospital beds, Table 1.3 reveals a negative treatment effect for nearly all treated regions, which is expected, given that reducing hospital bed numbers was one of the targets of the policy. Specifically, the bounds for $\tau_{g,r,t}$ are negative for Abruzzo (in the last two years), Calabria, Lazio (from 2009 onward), Molise (from 2008 onward), Sardegna (from 2010 onward), Sicilia (from 2007 onward), Piemonte, and Puglia (only in 2017 and 2018). In contrast, Liguria sometimes shows a positive effect, which is likely because Liguria exited the Recovery Plan.

As for the hospitalization rate, the bounds for the treatment effects are generally negative for nearly every year and region, indicating that the treatment effect was negative. However, for Liguria and Piemonte, the bounds cover zero in almost all post-treatment years, possibly due to these regions leaving the treatment status over time. Conversely, for Lazio, the upper bound is consistently lower than the lower bound in most post-treatment years. A similar pattern is observed for the hospitalization rate based on acute hospitalizations only, which is reported in Table F.1.

In Table 1.4, bounds for the ratio between the hospitalization for DRG at high risk of inappropriateness versus those that could not be avoided are presented. Except for Abruzzo (from 2010) and

Table 1.3: Bounds on treatment effect based on ($\delta_{T,Max}$, $\delta_{SC,Max}$)

	Abruzzo		**Calabria**		**Campania**		**Lazio**		**Liguria**		**Molise**		**Piemonte**		**Puglia**		**Sardegna**		**Sicilia**	
	ln (**current health exp.**)																			
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2007	-0.15	0.18			-0.13	0.07	-0.18	0.22	-0.11	0.09	-0.21	0.19					-0.15	0.05	-0.14	0.06
2008	-0.14	0.15			-0.12	0.05	-0.20	0.20	-0.12	0.08	-0.20	0.19					-0.10	0.07	-0.19	0.01
2009	-0.14	0.11			-0.10	0.06	-0.21	0.19	-0.10	0.04	-0.18	0.19					-0.05	0.11	-0.19	0.01
2010	-0.14	0.09	-0.12	0.08	-0.12	0.02	-0.25	0.15	-0.11	0.04	-0.18	0.17	-0.13	0.07			-0.03	0.11	-0.19	-0.00
2011	-0.16	0.07	-0.15	0.05	-0.14	-0.00	-0.27	0.12	-0.11	0.03	-0.20	0.15	-0.14	0.06	-0.12	0.08	-0.01	0.13	-0.19	-0.01
2012	-0.14	0.11	-0.12	0.08	-0.15	0.01	-0.27	0.12	-0.14	0.00	-0.18	0.18	-0.16	0.04	-0.11	0.09	0.00	0.16	-0.18	0.02
2013	-0.15	0.11	-0.13	0.07	-0.16	-0.00	-0.29	0.09	-0.15	0.02	-0.13	0.24	-0.17	0.03	-0.10	0.10	-0.01	0.15	-0.17	0.03
2014	-0.13	0.12	-0.12	0.08	-0.14	0.02	-0.29	0.07	-0.13	0.01	-0.18	0.19	-0.16	0.04	-0.09	0.11	0.01	0.16	-0.16	0.04
2015	-0.14	0.10	-0.13	0.07	-0.13	0.02	-0.28	0.07	-0.13	0.00	-0.21	0.15	-0.17	0.03	-0.09	0.11	0.01	0.16	-0.16	0.04
2016	-0.11	0.12	-0.11	0.09	-0.12	0.03	-0.29	0.07	-0.13	-0.01	-0.18	0.17	-0.17	0.03	-0.07	0.13	0.02	0.17	-0.14	0.06
2017	-0.09	0.12	-0.14	0.06	-0.10	0.01	-0.29	0.04	-0.12	-0.02	-0.20	0.12	-0.19	0.01	-0.10	0.10	0.00	0.11	-0.13	0.05
2018	-0.09	0.13	-0.11	0.09	-0.09	0.03	-0.28	0.02	-0.11	-0.01	-0.21	0.11	-0.19	0.01	-0.07	0.13	0.02	0.13	-0.11	0.08
	Hospital beds (ordinary)																			
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2007	-910.81	1560.00			-6209.00	4660.44	-2444.49	614.51	96.98	1816.00	-96.14	136.00					-1395.00	107.64	-2593.00	252.76
2008	-1385.66	948.00			-6544.30	4350.50	-1966.02	1092.98	-18.02	1414.00	-165.18	-14.00					-1440.00	125.64	-3307.00	-345.87
2009	-1536.12	696.00			-7255.56	3639.24	-4809.32	-1750.32	727.98	1651.00	-236.15	-79.00					-1454.00	122.76	-4162.00	-1225.23
2010	-1594.93	656.00	-1908.00	-156.04	-7451.37	3443.43	-5470.65	-2411.65	620.98	1543.00	-298.38	-139.00	-465.48	1545.52			-1807.00	-241.05	-3767.00	-816.71
2011	-1852.51	363.00	-2201.00	-457.04	-8068.32	2826.48	-7659.58	-4600.58	122.98	850.00	-490.94	-311.00	-731.09	1279.91	-3517.00	1476.68	-1800.00	-226.42	-4284.00	-1306.51

(Continues)

Table 1.3 (Continued)

	Abruzzo		Calabria		Campania		Lazio		Liguria		Molise		Piemonte		Puglia		Sardegna		Sicilia	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2012	-1976.44	249.00	-2862.00	-1092.04	-8584.75	2310.05	-7190.20	-4131.20	555.98	1378.00	-542.64	-380.00	-975.75	1035.25	-4473.00	545.02	-2301.00	-710.43	-4566.00	-1474.76
2013	-1664.41	364.00	-3409.00	-1553.04	-8334.59	2560.21	-7586.29	-4527.29	847.98	979.00	-555.82	-385.00	-611.98	1095.00	-4804.00	296.29	-2431.00	-774.57	-4708.00	-1455.22
2014	-1778.03	157.00	-3939.00	-2055.04	-8729.62	2165.18	-7510.01	-4451.01			-669.19	-537.00	-729.64	754.00	-5485.00	-356.41	-2786.00	-1099.98	-4996.00	-1717.19
2015	-1870.76	97.00	-4109.00	-2220.04	-8081.76	2813.04	-6694.14	-3635.14	1046.98	1082.00	-508.23	-382.00	-1104.23	497.00	-5058.00	75.73	-2681.00	-989.18	-5220.00	-1900.00
2016	-1738.25	156.00	-3840.00	-1947.04	-8565.13	2329.67	-6626.87	-3567.87			-501.96	-372.00	-1280.66	102.00	-5012.00	124.02	-2787.00	-1100.46	-5460.00	-2113.65
2017	-1857.37	-38.00	-3338.00	-1415.04	-8619.24	2275.56	-5441.67	-2382.67			-528.53	-445.00	-1536.67	-294.00	-5298.00	-154.00	-2955.00	-1250.12	-5289.00	-1887.91
2018	-1799.57	-16.00	-3443.00	-1456.04	-8836.59	2058.21	-6294.92	-3264.00			-637.07	-576.00	-1333.73	-279.00	-5243.00	-99.00	-2869.00	-1153.00	-5290.00	-1761.80

Hospitalization rate

	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2007	-22.63	17.17			-5.07	14.73	-12.62	-4.49	-22.37	19.43	-24.68	4.48					-23.23	11.77	-27.68	9.52
2008	-41.28	-2.76			-5.34	13.74	-12.23	-6.11	-23.05	18.75	-34.72	-6.38					-25.05	9.12	-33.70	3.50
2009	-60.09	-17.76			-7.36	9.11	-15.34	-11.77	-34.20	7.60	-48.66	-16.64					-32.83	-1.43	-45.73	-8.53
2010	-67.33	-25.78	-21.34	-1.23	-16.65	-2.31			-37.21	4.59	-41.12	-9.29	-2.85	9.55			-30.25	-1.03	-44.91	-8.26
2011	-68.53	-31.11	-29.17	-3.90	-17.83	-7.21			-49.79	-7.99	-55.36	-26.33	-4.32	8.08	-23.53	1.20	-27.17	-1.56	-47.75	-15.70
2012	-69.87	-32.59	-36.65	-12.38	-20.57	-10.12			-36.07	5.73	-67.06	-37.95	-5.64	6.76	-30.97	-6.77	-30.62	-5.28	-49.60	-16.97
2013	-67.89	-33.34	-44.81	-19.68	-23.52	-14.40			-33.95	7.85	-65.05	-37.66	-6.37	4.01	-34.88	-9.63	-35.36	-11.38	-54.14	-22.67
2014	-72.28	-38.08	-52.60	-27.96	-25.98	-16.69			-36.83	4.97	-83.07	-55.34	-8.97	0.17	-46.66	-21.78	-36.41	-12.29	-59.56	-28.04
2015	-72.93	-39.02	-56.81	-32.89	-23.70	-14.19			-36.72	5.08	-68.80	-41.75	-9.45	0.44	-39.94	-15.58	-37.56	-13.26	-62.04	-29.97
2016	-69.76	-38.28	-54.66	-28.36	-24.48	-18.03			-37.02	4.78	-69.90	-46.96	-11.82	-1.58	-41.90	-15.10	-38.41	-17.22	-62.68	-33.05
2017	-70.49	-40.34	-52.10	-23.61	-26.59	-22.00			-29.33	12.47	-71.90	-51.11	-14.05	-3.83	-44.42	-17.62	-39.04	-19.69	-60.92	-33.36
2018	-69.97	-41.26	-52.93	-22.78	-31.24	-28.73			-28.69	13.11	-76.52	-58.13	-13.36	-3.01	-43.11	-16.31	-10.21	7.03	-60.42	-34.67

Calabria (from 2014), no significant improvements in efficiency are observed in the treated regions after the introduction of the policy. On the other hand, in Campania and Sardegna (2009-2011) and Molise (2009-2010), inefficiency levels seem to have worsened.

The discrepancy between the estimated impact reported in Table 1.4 and the one obtained using the TWM approach could stem from at least one of the key identifying assumptions required for the TWM to provide a consistent estimator of the ATT being seriously violated. Alternatively, it is possible that the very small effects observed in Figure 1.4 were primarily driven by efficiency gains in Calabria and Abruzzo. When considering the estimated coefficients obtained by applying (1.11), the null hypothesis of coefficient insignificance is never rejected (results available upon request).

Regarding the c-section rate, PdRs appear to have had a positive effect. As shown in Table 1.4, the treatment effect is positive in almost all cases, except in cases where the estimated upper bound is lower than the lower bound. This result could suggest that, after the introduction of RPs, the appropriateness of care provided deteriorated. However, since c-sections are more common in private hospitals (De Luca et al., 2021), this finding may reflect a perceived decline in the quality of healthcare services provided by public hospitals, rather than a true decrease in the appropriateness of care. Again, the divergence between the estimated treatment effect here and the one obtained using the TWM approach may be due to violations of the key assumptions needed for TWM to yield a consistent ATT estimator.

Turning to the quality indicators, Table 1.5 reports results for the mortality rate from ischemic heart diseases and the percentage of patients migrating to other regions for ordinary acute hospitalizations. Regarding the former, an increase in the mortality rate from ischemic heart diseases is observed in Abruzzo (from 2009 to 2017), Campania (for four years), Liguria (from 2011 to 2016), and Molise (almost every year). For these regions, both the lower and upper bounds are typically above zero. In contrast, no clear patterns are found for the other treated regions. A similar pattern emerges when considering the all-cause mortality rate. However, when looking at this latter indicator, the fraction of cases where the upper bound is lower than the lower bound is higher than the mortality rate from ischemic heart diseases. The results for all-cause mortality are shown in Table F.1.

The absence of an effect of RPs on the mortality rate from ischemic heart diseases in Calabria could be due to an identification issue rather than the lack of effect. Calabria has consistently had the lowest mortality rate from ischemic heart diseases. To investigate this further, Figure 1.5 shows the time-series for the mortality rate from ischemic heart diseases in Calabria, Abruzzo and the average for the never-treated regions. From this chart, it is clear that the slight increase observed in Calabria following the introduction of the RP may have been offset by the significant decrease experienced by the never-treated regions. On the other hand, for Abruzzo, the mortality rate continued to rise after 2007 (when it first signed the RP), with the gap between its trend and that of the never-treated regions widening over time.

Regarding the impact of the policy on the percentage of patients seeking care outside their region of origin, Table 1.4 shows that after the introduction of the policy, a positive effect on mobility is detected for all treated regions except for Piemonte. However, this result should be interpreted with caution. By focusing solely on the percentage of patients, the spatial dimension of the process is completely overlooked. Additionally, the percentage is calculated across all regions, yet it is well-known that the flow of patients is not bidirectional in Italy (Balia et al., 2018; Arcà et al., 2020). Most of the flow comes from the Southern and, simultaneously, poorer regions of Italy, seeking care in hospitals in the Center and North of the country.

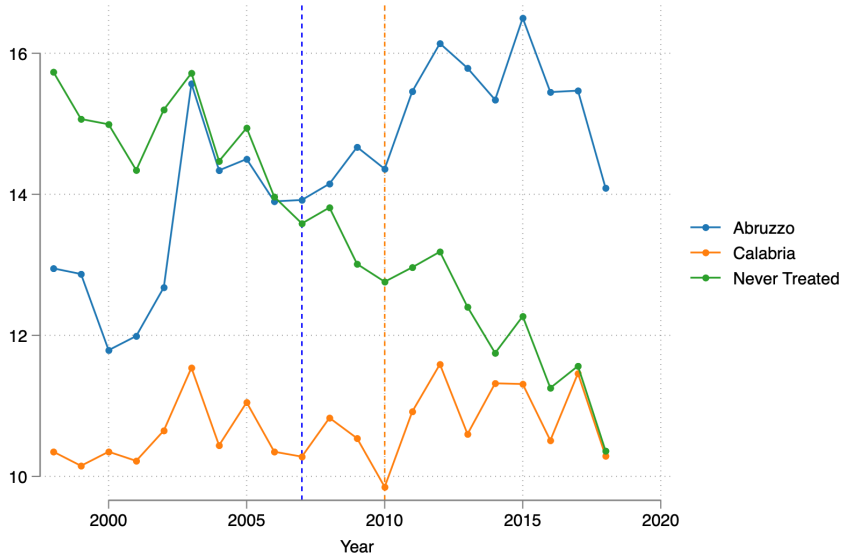
Table 1.4: Bounds on treatment effect based on ($\delta_{T,Max}$, $\delta_{SC,Max}$)

	Abruzzo		**Calabria**		**Campania**		**Lazio**		**Liguria**		**Molise**		**Piemonte**		**Puglia**		**Sardegna**		**Sicilia**	
	Inefficiency ratio																			
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2007	-0.14	0.06			-0.05	0.08	-0.12	0.08	-0.11	0.09	-0.14	0.06					-0.13	0.06	-0.10	0.10
2008	-0.18	0.02			-0.07	0.07	-0.13	0.07	-0.11	0.09	-0.15	0.05					-0.14	0.05	-0.13	0.07
2009	-0.06	0.06			0.15	0.23	0.01	0.07	-0.04	0.07	0.13	0.19					0.10	0.12	-0.09	0.05
2010	-0.15	-0.01	-0.09	0.11	0.09	0.20	-0.04	0.05	-0.04	0.07	0.02	0.11	-0.09	0.11			0.08	0.13	-0.13	0.04
2011	-0.17	-0.03	-0.14	0.06	0.04	0.13	-0.06	0.03	-0.06	0.06	-0.06	0.02	-0.10	0.10	-0.12	0.28	0.06	0.11	-0.15	0.01
2012	-0.25	-0.05	-0.17	0.03	-0.07	0.12	-0.16	0.04	-0.13	0.06	-0.16	0.02	-0.09	0.11	-0.16	0.22	-0.07	0.08	-0.17	0.03
2013	-0.26	-0.06	-0.19	0.01	-0.08	0.11	-0.17	0.03	-0.13	0.07	-0.17	0.02	-0.09	0.11	-0.17	0.21	-0.09	0.06	-0.19	0.01
2014	-0.27	-0.07	-0.21	-0.01	-0.08	0.12	-0.17	0.03	-0.11	0.09	-0.19	0.01	-0.09	0.11	-0.19	0.18	-0.11	0.05	-0.19	0.01
2015	-0.25	-0.05	-0.22	-0.02	-0.07	0.10	-0.16	0.04	-0.13	0.07	-0.18	0.02	-0.09	0.11	-0.21	0.13	-0.14	0.06	-0.19	0.01
2016	-0.25	-0.05	-0.23	-0.03	-0.08	0.09	-0.15	0.05	-0.14	0.06	-0.18	0.02	-0.10	0.10	-0.23	0.11	-0.15	0.05	-0.19	0.01
2017	-0.24	-0.04	-0.21	-0.01	-0.08	0.07	-0.16	0.04	-0.10	0.10	-0.18	0.02	-0.12	0.08	-0.23	0.09	-0.14	0.04	-0.17	0.03
2018	-0.25	-0.05	-0.22	-0.02	-0.12	0.03	-0.15	0.05	-0.09	0.11	-0.20	-0.00	-0.12	0.08	-0.24	0.08	-0.15	0.03	-0.16	0.04
	% c-sections																			
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2007	1.56	5.25			-0.58	5.54	1.00	3.80	1.00	3.21	1.44	7.46					-0.59	4.40	2.79	4.66
2008	2.99	6.50			0.08	6.46	-0.24	2.56	-0.24	2.95	1.58	6.82					-1.45	3.78	2.78	4.31
2009	2.51	5.62			1.75	6.94	0.63	3.43	0.96	3.87	0.30	5.31					-0.63	3.42	4.08	4.82
2010	2.07	5.31	-3.75	1.56	1.48	7.02	0.97	3.77	1.82	4.74	0.79	6.44	-1.42	1.57			-0.21	4.18	4.22	5.09
2011	3.25	5.64	0.17	4.22	2.31	6.81	0.80	3.60	2.34	5.01	-1.94	2.89	-2.37	-0.07	-0.54	3.18	1.89	5.24		
2012	2.11	2.44	-2.30	-1.30	6.06	7.46	1.21	3.63	1.38	3.01	0.29	3.04	-2.25	0.07	1.84	2.51	5.10	5.35		
2013							1.01	3.19	0.47	2.82	2.32	3.39	-1.05	0.03						
2014							0.88	2.52	1.52	3.18	3.32	5.31	-2.70	-0.54						
2015							0.13	1.88	-0.53	2.35	0.71	1.68	-2.80	-1.74						
2016							-1.19	-0.02	-0.56	1.87			-3.35	-2.07						
2017							-1.88	-0.49	-3.34	-0.14			-3.53	-2.22						
2018							-1.08	-0.49	-4.15	-1.45			-3.98	-3.43						

Table 1.5: Bounds on treatment effect based on ($\delta_{T,Max}$, $\delta_{SC,Max}$)

	Abruzzo		Calabria		Campania		Lazio		Liguria		Molise		Piemonte		Puglia		Sardegna		Sicilia	
	Mortality rate from ischaemic heart diseases																			
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2007	-0.84	0.36			-0.57	0.43	-0.72	0.08	-1.43	-0.03	-0.72	1.48					-0.10	0.10	-0.25	0.75
2008	-0.41	0.79			-0.31	0.69	-0.45	0.35	-0.59	0.81	0.82	3.02					0.19	0.39	-0.02	0.98
2009	0.78	1.97			0.51	1.42	-0.82	-0.02	-0.38	1.02	-1.25	0.74							0.72	1.25
2010	0.40	1.60	-1.32	0.28	0.21	1.21	-0.30	0.50	0.00	1.40	0.12	2.32	0.27	0.55					-0.04	0.87
2011	0.92	2.12	-0.72	0.52	0.20	1.03	0.27	1.07	0.85	2.25	1.52	3.72	0.27	0.80	-0.47	1.08	0.54	0.74	-0.01	0.99
2012	1.04	2.09					-0.46	0.34	1.71	3.11	-0.10	2.10	0.88	1.36	-0.09	0.04	-0.52	-0.32		
2013	1.39	2.59	-1.04	-0.34	-0.10	0.19	-0.26	0.54	1.52	2.92	0.91	3.11			-0.67	0.41			-0.75	0.13
2014	1.83	2.64	-0.32	1.13	-0.31	0.69	-0.46	-0.00	1.43	1.91	2.46	4.40			-0.67	1.13			-0.52	0.48
2015	2.40	3.60	-0.33	0.57	0.86	1.35	-0.36	0.44	0.59	1.70	3.18	5.38			0.07	1.39			-0.63	0.37
2016	1.86	2.75	-1.13	-0.23	-0.40	0.09	-0.53	-0.45	1.11	1.24	-0.14	1.70			-1.20	0.26			-1.45	-0.45
2017	1.81	2.77	-0.18	0.44	-0.33	-0.12	-0.48	-0.25			1.69	3.89			-0.37	0.78			-1.27	-0.27
2018			-0.82	0.78	-0.34	0.66					2.33	2.52			-0.20	1.60			-0.36	-0.18
	% Patients migrating for ordinary acute hospit.																			
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2007	0.81	2.12			-0.37	0.45	-0.41	0.45	-0.13	1.47	-2.84	-1.15					-0.28	1.20	-0.73	0.63
2008	3.19	4.26			-0.70	0.30	-0.69	0.23	0.19	1.78	-2.45	-0.51					-0.09	1.21	-0.58	1.10
2009	5.33	5.73			0.12	0.55	-0.63	0.36	0.38	1.96	-1.14	0.86					-0.00	1.11	0.21	1.20
2010	6.25	6.59	-0.36	0.88	0.54	1.05	-0.43	0.50	1.07	2.57	-2.06	-0.06	-1.21	-0.24			-0.16	1.07	0.25	1.32
2011			0.79	1.10			0.18	1.08	1.80	3.31	-0.15	1.85	-1.22	-0.28	0.91	0.99	0.36	1.30	1.35	1.55
2012	5.30	5.80	0.56	2.15	0.37	1.11	0.21	1.15	2.17	3.50	0.70	2.70	-1.15	0.09	0.08	1.28	-0.04	1.27	-0.22	1.17
2013	5.34	5.79	1.78	3.14	0.54	1.08	0.63	1.59	1.99	3.47	1.64	3.64	-0.95	0.28	0.35	1.48	0.02	1.23	0.02	1.20
2014			3.32	3.96			1.18	2.15	2.36	3.76	2.78	4.78	-0.98	0.27	1.33	1.74	0.11	1.07	0.78	1.29
2015	5.69	5.94	3.49	4.51	1.23	1.49	1.66	2.55	2.36	3.96	2.85	4.85	-0.83	-0.07	1.26	2.05	0.21	1.30	0.67	1.54
2016	5.12	5.99	3.20	5.07	0.81	1.81	2.02	2.67	3.75	5.35	4.68	6.68	-0.31	0.31	1.16	2.36	0.10	1.67	0.26	1.93
2017	4.75	5.73	2.66	4.45	1.29	2.28	1.77	2.48	2.17	3.77	5.17	7.10	-0.68	-0.11	1.38	2.58	-0.50	0.96	0.45	2.06
2018	5.15	6.33	2.21	4.30	1.58	2.58	1.76	2.35	2.08	3.68	6.01	7.73	-0.82	-0.37	1.51	2.71	0.84	2.44	0.38	2.09

Figure 1.5: Time Series Evolution – Mortality rate from ischaemic heart diseases



Notes: The blue dotted line represents the year Abruzzo signed a PdR. Whereas the orange line represents the first year in which Calabria was treated.

As highlighted by Arcà et al. (2020), the detection of an impact could suggest that the introduction of the policy has worsened the gap in equality of access to healthcare between the rich and the poor, while simultaneously amplifying the existing divide between Southern and Center-Northern Italy in terms of the quality of services provided.

Lastly, in Table F.1, the results for the percentage of patients with hip fracture replacements are presented. What stands out is that the upper bound is consistently lower than the lower bound in almost every post-treatment year and for each region, which means that the treatment effect is not identified. There are two potential explanations for this lack of effect on this indicator. First, it is essential to note that this indicator is part of the ex-post monitoring process. Consequently, both hospitals and regions have economic incentives to closely monitor this indicator and ensure that adequate levels of this target are met. Second, the absence of a detected effect may also be due to an identification issue, as evidenced by the fact that, in most cases, the upper bound is lower than the lower bound.

1.7.3 Robustness checks

In this subsection, the results obtained by carrying out the different sensitivity exercises discussed in Section 1.6.4 will be presented.

Figure E.2 of Appendix E plots $\hat{\tau}_s$ obtained by estimating (1.11) using as dependent variables the log of current health expenditure (Panel (a)) and the mortality rate from ischaemic heart diseases (Panel (b)). Coefficients are reported with 95% bootstrap confidence intervals obtained via subcluster wild bootstrap. Along the y-axis, point estimates are shown. On the x-axis, the estimated coefficients for $w_{g,t}fs_t$ are reported.

The results presented in Figure E.2 in Appendix E are consistent with those shown in Figures 1.3 and 1.4. The estimated effect on the mortality rate is always positive and statistically significant, while the impact on current health expenditure remains negative. The only difference compared to the results

in Figures 1.3 and 1.4 is that the first three estimated coefficients (for 2007, 2008, and 2009) are now statistically indistinguishable from zero. This lack of statistical significance for these coefficients may be because only seven regions were treated during these three years. In contrast, Calabria and Piemonte joined the PdR in 2010, followed by Puglia in 2011.[32]

To further assess the robustness of the results obtained via the TWM approach, the estimator proposed by Sun and Abraham (2021) was also used. The results obtained using this estimator are similar to those obtained via the TWM. For brevity, only the results for the log of current health expenditure are presented in Figure E.3, with results for the other indicators available upon request. From Figure E.3, it is clear that the point estimates are comparable to those obtained by estimating (1.5). However, two key points should be highlighted. First, as mentioned earlier, the estimator proposed by Sun and Abraham (2021) does not allow for relaxing the Parallel Trends and No Anticipation assumptions. Second, the estimated coefficients in Figure E.3 are accompanied by 95% confidence intervals, where standard errors were computed using a standard CRVE, with clustering at the RHS level.[33] Overall, due to these limitations, we prefer the results obtained using the TWM approach.

To assess the robustness of the results obtained via bounds, these were re-estimated by jointly imposing the two δ's to be equal to the 75^{th} percentile of the absolute difference observed in the pre-treatment period (p_{75}). The results based on $\delta_{SC,p_{75}}$ and $\delta_{time,p_{75}}$ are presented in Table F.2. For brevity, only the results for the mortality rate from AMI are shown (results for the other indicators are available upon request). From Table F.2, it is evident that while the fraction of bounds not identified is slightly larger compared to those in Table 1.5, the results remain consistent. Overall, it appears that after the introduction of RPs, there was an increase in the mortality rate from ischaemic heart diseases.

Furthermore, as part of a further sensitivity exercise, a complete set of combinations of δ_{SC} and δ_{time} was also used. To save space, only the results for the mortality rate from ischaemic heart diseases in Abruzzo are shown in Table F.4. These results align with those presented in Table 1.5 for Abruzzo (particularly from 2014 onward), and hold across different combinations of δ_{SC} and δ_{time}. Interestingly, when both δ_{SC} and δ_{time} are set to 0, the upper bound is consistently smaller than the lower bound, suggesting that the time invariance and SC assumptions do not hold jointly in the data. This finding aligns with the results from Depalo (2019). Similar findings were obtained for other regions using different combinations of δ_{SC} and δ_{time}, confirming the robustness of the results obtained via the bounds approach.

Several sensitivity exercises were performed to assess whether the (C)NA assumption is likely to be satisfied. First, equations (1.10) and (1.11) were re-estimated by including up to three leads of the treatment dummy to test for the presence of anticipatory effects. For all the dependent variables considered, I consistently fail to reject the null hypothesis of joint insignificance for the coefficients of the leads, except for the c-section rate. These results are available upon request.

Second, equations (1.2) through (1.11) were re-estimated by anticipating each region's entry into the PdR by two years. For brevity, these results are not shown. What emerges is that the coefficients

[32] When the ATT is allowed to vary by calendar time, as in equations (1.11), the interpretation of the coefficients changes. Rather than reflecting the effect of treatment based on how many periods have passed since the region was first treated, the coefficients $\hat{\tau}_s$ represent the effect for regions treated in a specific year $t = s$. For example, $\hat{\tau}_{2015}$ will reflect how much being treated in 2015 affects the dependent variable. Results from estimating (1.11) with other indicators as the dependent variable are available upon request.

[33] The STATA package `eventstudyinteract` provided by Sun and Abraham (2021) does not permit the use of bootstrap.

for $w_{g,t}intens(0)$, $w_{g,t}intens(1)$, $w_{g,t}f_{2005}$, and $w_{g,t}f_{2006}$ are all statistically indistinguishable from zero, indicating no evidence of anticipatory effects. This holds for all dependent variables, except for the hospitalization rate, where the coefficients are positive and statistically significant.

Next, results obtained by re-estimating equations (1.12) and (1.13) while excluding later-treated regions align perfectly with those presented in Subsection 1.7.1. These results are also available upon request.

Overall, these findings suggest that, except for the c-section rate, anticipatory behaviors are unlikely to be a significant issue in this context.

Lastly, Figure E.4 presents the results obtained by estimating (1.5), using as dependent variables the log of current health expenditure and the mortality rate from AMI, while excluding leavers from the analysis. We observe that, apart from the confidence intervals for τ_{-4} and τ_{-3} in Panel (b), which are now less precisely estimated, the results remain consistent with those reported in Figure 1.3 and Figure 1.4.

Furthermore, the test of pre-trends aligns with those presented in Table E.1 (not shown): we still reject H_0 for the log of current health expenditure, whereas we fail to reject H_0 when using the mortality rate from AMI as the dependent variable. Results for the other indicators are also consistent with those in the main text and are available upon request.

Overall, while the bounds suggest that regions exiting treatment status may have violated the irreversibility of treatment assumption, the TWM estimates do not appear to be affected. This indicates that, even if such a violation occurs, it represents only a minor issue in this context.

1.7.4 Discussion

While this study relies on administrative data–which are generally less prone to measurement errors–the available data have several limitations, as explained in Section 1.5. Since PdRs are region-specific plans, all treated regions were required to implement budget cuts to restore financial balance by cutting inefficient spending. In practice, however, these cuts often translate into an accounting exercise, with spending reductions applied roughly linearly. Regions–and, more specifically, hospitals–retain some flexibility in determining how to achieve these cost reductions (e.g., through staff reductions or hospital network reorganizations), provided that ELAs are maintained. However, due to the aggregated nature of the data, capturing this heterogeneity remains challenging, representing a limitation of this study.

Another limitation is that, although multiple cost indicators are analyzed, not all mechanisms through which PdRs may have impacted health outcomes are directly observable. Beyond reducing staffed beds and hospitalization rates, regions under commissioners were also required to implement stricter measures. These measures led to reduced hospital and NHS personnel (e.g., Bordignon et al., 2020), as well as longer travel times due to hospital closures (Ghislandi et al., 2025). These changes often result in longer waiting lists, higher bed occupancy rates, lower admission rates, and increased readmission rates (Blom et al., 2014). While PdRs may have directly affected indicators subject to ex-post monitoring, other outcomes considered in this study were likely influenced indirectly through these channels. The effects, however, may vary between and within regions depending on the strength of these measures.

The inability to observe all potential channels through which PdRs may have affected health outcomes, combined with the lack of more detailed data, makes it challenging to pinpoint the exact mech-

anisms behind the previously documented deterioration in health outcomes. Moreover, this limitation restricts the ability to assess the broader impact of PdRs on patient health beyond mortality rates, including survival following hospital procedures, hospital readmissions, post-surgical complications, and considering patients' existing comorbidities.

Furthermore, even after including controls such as population size and the age and gender distribution to standardize mortality rates, these adjustments may not be sufficient in the case of acute or emergency events. This is particularly relevant for mortality from AMI, where the probability of emergency occurrences is relatively high. Consequently, the risk adjustment is incomplete and we may fail to fully capture territorial differences–such as variations in exposure to pollutants–in mortality rates among individuals of the same gender and age across regions.

Acknowledging these limitations and calling for future research in these areas, there is evidence that, following the introduction of PdRs, the Central Government effectively reduced costs and current health expenditure in treated regions. Although I cannot directly observe all the potential channels through which these cost reductions (e.g., reduction in hospital personnel) were achieved, I find evidence of a decrease in hospital beds and hospitalization rates, suggesting that cost containment was likely achieved through these two targets as well. As noted by Depalo (2019), reducing hospitalization rates is the most straightforward way to cut healthcare spending, as they account for approximately 90% of the current health expenditure of Italian RHSs.

Regarding current health expenditure, as explained in the previous subsections, the lack of an effect when considering bounds could be due to identification issues rather than an actual absence of impact. This hypothesis is supported by the substantial reduction in the total deficit for public health spending documented after 2007, as shown in Figure 1.1.

The Central Government attempted, with the introduction of PdRs, to reduce costs and restore budget balance by cutting inefficient spending components. One proxy for inefficient resource use is the ratio between the hospitalization for DRG at high-risk inappropriateness versus those that cannot be avoided. What emerges is that, except Abruzzo and Calabria, the introduction of RPs did not reduce the level of inefficiencies in treated RHSs. Furthermore, a positive effect is observed when considering the percentage of c-sections as a proxy for care inappropriateness. However, this result may be due to an increased demand for c-sections in private hospitals (De Luca et al., 2021), rather than public hospitals performing more c-sections.

On the other hand, a deterioration in the quality of services provided (proxied by the mortality rate from ischaemic heart diseases) and in the perceived quality (proxied by the percentage of patients migrating to other regions for acute hospitalizations) is observed in treated regions, with these results remaining statistically significant in the long run. Furthermore, introducing RPs seems to have had a negative–but somewhat smaller–effect on the mortality rate (from all causes) in treated RHSs.

The fact that all of these indicators appear to have been affected by the policy is consistent with existing studies (e.g. Brown et al., 2014; Kwok et al., 2018). Specifically, regarding the mortality rate from AMI, PdRs may have negatively impacted this indicator in several ways. First, AMI is a highly costly procedure (Kwok et al., 2018; Lobo et al., 2020), meaning healthcare spending cuts could have affected this indicator through reduced hospital budget. Additionally, reduced hospital supply has been linked to higher 30-day readmission rates after AMI (Brown et al., 2014) and increased mortality from AMI (Borra et al., 2020). Indeed, reducing hospital beds often leads to higher bed occupancy rates, which, in turn, can increase readmission rates and are associated with higher mortality (Blom et al., 2014; Madsen

et al., 2014; Boden et al., 2016). Specifically in the Italian context, Arcà et al. (2020) documents a rise in avoidable mortality following a decrease in hospital supply, while Ghislandi et al. (2025) finds an increase in in-hospital mortality after hospital closures, attributing this effect to the longer travel times required to reach the nearest hospital.

Although the exact channels through which this indicator was affected cannot be precisely determined, existing literature suggests that the observed increase in mortality from AMI may stem from at least one of the factors discussed above.

Likewise, the observed effect of the policy on patient mobility is not surprising, as hospital supply is one of the main predictors of interregional patient mobility in Italy (Balia et al., 2018). These findings align with previous studies highlighting that generalized spending cuts may have adverse health effects.

Furthermore, the consistency of the ATT estimates derived from the TWM approach with the treatment effects obtained using the non-parametric estimator proposed by Manski and Pepper (2018) for almost all the indicators examined (except for the log of current health expenditure, the c-section rate, and the inefficiency ratio), along with their robustness across various sensitivity analyses, suggests that there are no significant violations of the key identifying assumptions necessary for the consistency of the parametric estimator.

Although I cannot entirely rule out the possibility that the policy may have impacted the indicators considered in this study through other channels not explicitly observed, I can investigate whether the largest increase in these outcomes can be partially attributed to reduced hospitalization rates and hospital beds. To this end, rank correlation coefficients were employed to examine whether regions experiencing larger reductions in hospital beds and hospitalization rates also saw a greater increase in the mortality rate from ischaemic heart diseases and patient mobility. Precisely, Kendall's τ and Spearman's ρ were calculated.

Both correlation coefficients indicate a negative association between the bounds for $\tau_{g,r,t}$ for the hospitalization rate and the bounds for the mortality rate from ischaemic heart diseases. Similarly, the rank correlation between the bounds for the treatment effect for hospitalization rates and that for patients' mobility is also negative. The larger the reduction in hospitalization rates, the worse the consequences on quality indicators, which is consistent with existing studies in medicine and health economics that consider hospitalization rates as a proxy for quality rather than costs (e.g., Berchialla et al., 2010). By contrast, the null hypothesis of independence between the bounds for hospital beds and those for quality indicators is never rejected.[34]

The fact that regions with the largest reductions in hospitalization rates also experienced the most significant negative consequences on the quality of their RHS aligns with the findings of Depalo (2019). This heterogeneity may be explained by the stricter measures imposed in some regions.

Overall, the findings of this study are consistent with both national and international research, which shows that budget cuts in healthcare–achieved through reductions in hospitalization rates, hospital beds, and hospital personnel–do not necessarily lead to increased efficiency among healthcare providers (e.g., McKay and Deily, 2008; Hunt and Link, 2020). On the contrary, such cuts are often associated with poorer health outcomes (Heijink et al., 2013; Siverskog and Henriksson, 2022; Arcà et al., 2020). This supports the idea that generalized cost reductions, rather than cutting inefficient use of

[34] Although the specific channels through which PdRs may have influenced the outcomes in this study are not directly observed, the use of bounds, as explained earlier, enables the estimation of region-specific effects, partially accounting for regional variations.

resources, may have detrimental effects on health outcomes (McKay and Deily, 2008). Additionally, these results align with existing literature on PdRs (e.g., Depalo, 2019; Bordignon et al., 2020; Cirulli and Marini, 2023; Beraldo et al., 2023).

1.8 Conclusions

Healthcare spending cuts are a common strategy to restore budget balance, particularly in countries with publicly funded NHS systems. However, assessing their impact on health outcomes remains a key concern for policymakers.

This paper contributes to the ongoing debate by focusing on Italy. The Italian NHS is particularly interesting to analyze because it is a decentralized, almost entirely publicly funded system in which regions retain RHS's management power. Over the years, regions have gradually been granted more responsibilities in managing RHSs to contain costs. However, despite these efforts, the Central Government has consistently provided ex-post funding to cover the large deficits run by regional governments. This practice has led to a dramatic increase in public expenditure and the total deficit for public health spending. Since 2007, the Central Government has introduced region-specific recovery plans to curb excessive spending and restore the budget balance of RHSs while ensuring the provision of Essential Levels of Assistance.

While there is broad consensus that this policy effectively reduced costs, existing studies find contrasting evidence regarding its impact on the quality and efficiency of RHSs.

The primary goal of this study is to provide novel evidence on Piani di Rientro's impact on the quality and efficiency of RHSs. To this end, in addition to indicators part of the central government's ex-post monitoring, other commonly-used quality proxies are considered. This approach allows me to test the hypothesis whether regions have strategically outperformed on indicators that discriminate on their ability to receive funds.

To estimate the effect of the policy on these indicators, unlike previous studies on PdRs, I directly account for the staggered nature of the policy by adopting the Two-Way Mundlak approach proposed by Wooldridge (2021). This method also allows me to assess the policy's long-run impact.

The results from the Two-Way Mundlak approach are complemented by those obtained under milder non-parametric assumptions, using the estimator proposed by Manski and Pepper (2018). Using bounds is crucial in this context, as the key assumptions required for consistency of the parametric estimator may not hold for some of the indicators considered. Additionally, bounds allows for identifying region-specific treatment effects, shedding further light on the findings from the parametric analysis.

Despite the limitations of the data, as discussed in the previous section, the findings indicate that Piani di Rientro effectively reduced costs, where cost containment was achieved, among other channels, by reduced hospital beds and hospitalization rates. However, when assessing the impact on inefficient use of resources, the results do not suggest significant efficiency gains, except in Abruzzo and Calabria. Moreover, the policy led to a deterioration in the quality of services in treated regions, with the most severe adverse effects observed in regions that experienced the sharpest reductions in hospitalization rates. These findings are robust across various sensitivity analyses and persist even in the long run. Further, for almost all variables, the results from the parametric approach align with those estimated using bounds.

In conclusion, these findings align with international research documenting that healthcare spending cuts do not always lead to efficiency gains and can instead result in worsened health outcomes. I believe these findings and the observation that regions implementing more gradual cost reductions experienced less severe declines in service quality may inform policy makers about the importance of taking a more gradual approach to healthcare spending cuts. In addition, broad, indiscriminate spending cuts, rather than targeted reductions addressing inefficiencies, may ultimately be detrimental to public health.

Bibliography

Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies*, 72(1):1–19.

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425.

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. *American economic review*, 93(1):113–132.

Aimone Gigio, L., Alampi, D., Camussi, S., Ciaccio, G., Guaitini, P., Lozzi, M., Mancini, A. L., Panicara, E., and Paolicelli, M. (2018). La sanità in Italia: il difficile equilibrio tra vincoli di bilancio e qualità dei servizi nelle regioni in piano di rientro. *Questioni di Economia e Finanza in Bank of Italy*, 427:1–100.

Amarilla-Donoso, F. J., López-Espuela, F., Roncero-Martín, R., Leal-Hernandez, O., Puerto-Parejo, L. M., Aliaga-Vera, I., Toribio-Felipe, R., and Lavado-García, J. M. (2020). Quality of life in elderly people after a hip fracture: a prospective study. *Health and quality of life outcomes*, 18:1–10.

Andrews, M., Elamin, O., Hall, A. R., Kyriakoulis, K., and Sutton, M. (2017). Inference in the presence of redundant moment conditions and the impact of government health expenditure on health outcomes in England. *Econometric Reviews*, 36(1-3):23–41.

Angelillo, I., Ricciardi, G., Nante, N., Boccia, A., Bianco, A., La Torre, G., Vinci, V., and De Giusti, M. (2000). Appropriateness of hospital utilisation in Italy. *Public health*, 114(1):9–14.

Arcà, E., Principe, F., and Van Doorslaer, E. (2020). Death by austerity? The impact of cost containment on avoidable mortality in Italy. *Health Economics*, 29(12):1500–1516.

Balia, S., Brau, R., and Marrocu, E. (2018). Interregional patient mobility in a decentralized healthcare system. *Regional Studies*, 52(3):388–402.

Bei, X. (2024). Inference on union bounds with applications to did, rdd, bunching, and structural counterfactuals.

Beraldo, S., Collaro, M., and Marino, I. (2023). Patient migration as a response to the regulation of subnational healthcare budgets. *Regional Studies*, 0(0):1–13.

Berchialla, P., Foltran, F., Di Lenarda, A., Barbati, G., Franchin, L., Salvatore, L., Gregori, D., et al. (2010). Hospitalization rate as a proxy for the quality of life evaluation: the case of dilated cardiomiopathy. *Journal of preventive medicine and hygiene*, 51(4):152–156.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1):249–275.

Blom, M. C., Jonsson, F., Landin-Olsson, M., and Ivarsson, K. (2014). The probability of patients being admitted from the emergency department is negatively correlated to in-hospital bed occupancy–a registry study. *International journal of emergency medicine*, 7:1–7.

Boden, D., Agarwal, A., Hussain, T., Martin, S., Radford, N., Riyat, M., So, K., Su, Y., Turvey, A., and Whale, C. (2016). Lowering levels of bed occupancy is associated with decreased inhospital mortality and improved performance on the 4-hour target in a UK District General Hospital. *Emergency Medicine Journal*, 33(2):85–90.

Bordignon, M., Coretti, S., Piacenza, M., and Turati, G. (2020). Hardening subnational budget constraints via administrative subordination: The Italian experience of recovery plans in regional health services. *Health Economics*, 29(11):1378–1399.

Borra, C., Pons-Pons, J., and Vilar-Rodriguez, M. (2020). Austerity, healthcare provision, and health outcomes in Spain. *The European Journal of Health Economics*, 21:409–423.

Borusyak, K., Jaravel, X., and Spiess, J. (2021). Revisiting event study designs: Robust and efficient estimation. *arXiv preprint:2108.12419*.

Brown, J. R., Chang, C.-H., Zhou, W., MacKenzie, T. A., Malenka, D. J., and Goodman, D. C. (2014). Health system characteristics and rates of readmission after acute myocardial infarction in the United States. *Journal of the American Heart Association*, 3(3):e000714.

Calabrò, R. (2016). Servono terapie innovative per le regioni in piano di rientro (in Italian). *Il Sole 24 ore, 4 Oct.*

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

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

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources*, 50(2):317–372.

Cirulli, V. and Marini, G. (2023). Are austerity measures really distressing? Evidence from Italy. *Economics & Human Biology*, page 101217.

Crémieux, P., Ouellette, P., and Pilon, C. (1999). Health care spending as determinants of health outcomes. *Health Economics*, 8(7):627–639.

Davidson, R., MacKinnon, J. G., et al. (2004). *Econometric theory and methods*, volume 5. Oxford University Press New York.

De Chaisemartin, C. and d'Haultfoeuille, X. (2022a). Difference-in-differences estimators of intertemporal treatment effects. Technical report, National Bureau of Economic Research.

De Chaisemartin, C. and d'Haultfoeuille, X. (2022b). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. Technical report, National Bureau of Economic Research.

De Chaisemartin, C. and d'Haultfoeuille, X. (2023). Two-way fixed effects and differences-in-differences estimators with several treatments. *Journal of Econometrics*, 236(2):105480.

De Luca, G., Lisi, D., Martorana, M., and Siciliani, L. (2021). Does higher institutional quality improve the appropriateness of healthcare provision? *Journal of Public Economics*, 194:104356.

Depalo, D. (2019). The side effects on health of a recovery plan in Italy: A nonparametric bounding approach. *Regional Science and Urban Economics*, 78:103466.

Eurostat (2013). European system of accounts - ESA 2010. Technical report, Eurostat.

Ferman, B. and Pinto, C. (2016). Revisiting the synthetic control estimator.

Franklin, B., Hochlaf, D., and Holley-Moore, G. (2017). Public Health in Europe during the austerity years. In *A research report from ICL-UK*. ICL-IK London.

Ghislandi, S., Renner, A.-T., and Varghese, N. (2025). The impact of budget cuts on individual patient health: Causal evidence from hospital closures. *Journal of Health Economics*, page 102975.

Giancotti, M., Sulku, S. N., Pipitone, V., Mauro, M., et al. (2020). Do Recovery Plans Improve Public Hospitals Efficiency and Productivity? Evidence from Italy. *International Review of Business Research Papers*, 16(1).

Golinelli, D., Toscano, F., Bucci, A., Lenzi, J., Fantini, M. P., Nante, N., and Messina, G. (2017). Health expenditure and all-cause mortality in the 'galaxy' of Italian regional healthcare systems: a 15-year panel data analysis. *Applied health economics and health policy*, 15:773–783.

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

Gregory, K. D., Jackson, S., Korst, L., and Fridman, M. (2012). Cesarean versus vaginal delivery: whose risks? Whose benefits? *American journal of perinatology*, 29(01):07–18.

Guccio, C., Pignataro, G., Romeo, D., and Vidoli, F. (2022). Is austerity good for efficiency, at least? A counterfactual assessment for the Italian NHS. Health, Econometrics and Data Group (HEDG) Working Papers 22/28, HEDG, c/o Department of Economics, University of York.

Guccio, C., Pignatora, G., and Vidoli, F. (2023). It never rains but it pours: Austerity and mortality rate in peripheral areas. Health, Econometrics and Data Group (HEDG) Working Papers 23/02, HEDG, c/o Department of Economics, University of York.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, 66(5):1017–1098.

Heckman, J. J., Humphries, J. E., and Veramendi, G. (2016). Dynamic treatment effects. *Journal of Econometrics*, 191(2):276–292. Innovations in Measurement in Economics and Econometrics.

Heijink, R., Koolman, X., and Westert, G. (2013). Spending more money, saving more lives? The relationship between avoidable mortality and healthcare spending in 14 countries. *The European Journal of Health Economics*, 14(3):527–538.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Hunt, D. J. and Link, C. R. (2020). Better outcomes at lower costs? the effect of public health expenditures on hospital efficiency. *Applied Economics*, 52(4):400–414.

Imai, K. and Kim, I. S. (2021). On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data. *Political Analysis*, 29(3):405–415.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86.

Institute of Medicine (2001). *Crossing the Quality Chasm: A New Health System for the 21st Century*. The National Academies Press, Washington, DC.

Kwok, C. S., Rao, S. V., Potts, J. E., Kontopantelis, E., Rashid, M., Kinnaird, T., Curzen, N., Nolan, J., Bagur, R., and Mamas, M. A. (2018). Burden of 30-day readmissions after percutaneous coronary intervention in 833,344 patients in the United States: predictors, causes, and cost: insights from the Nationwide Readmission Database. *JACC: Cardiovascular Interventions*, 11(7):665–674.

Leive, A., David, G., and Candon, M. (2023). On resource allocation in health care: The case of concierge medicine. *Journal of Health Economics*, page 102776.

Lobo, M. F., Azzone, V., Lopes, F., Freitas, A., Costa-Pereira, A., Normand, S.-L., and Teixeira-Pinto, A. (2020). Understanding the large heterogeneity in hospital readmissions and mortality for acute myocardial infarction. *Health Policy*, 124(7):684–694.

MacKinnon, J. G., Nielsen, M. Ø., and Webb, M. D. (2023). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics*, 232(2):272–299.

MacKinnon, J. G. and Webb, M. D. (2017). Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics*, 32(2):233–254.

MacKinnon, J. G. and Webb, M. D. (2018). The wild bootstrap for few (treated) clusters. *The Econometrics Journal*, 21(2):114–135.

Madsen, F., Ladelund, S., and Linneberg, A. (2014). High levels of bed occupancy associated with increased inpatient and thirty-day hospital mortality in Denmark. *Health affairs*, 33(7):1236–1244.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323.

Manski, C. F. (2003). *Partial identification of probability distributions*, volume 5. Springer.

Manski, C. F. and Pepper, J. V. (2013). Deterrence and the death penalty: Partial identification analysis using repeated cross sections. *Journal of Quantitative Criminology*, 29:123–141.

Manski, C. F. and Pepper, J. V. (2018). How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions. *The Review of Economics and Statistics*, 100(2):232–244.

Martini, G., Berta, P., Mullahy, J., and Vittadini, G. (2014). The effectiveness–efficiency trade-off in health care: The case of hospitals in Lombardy, Italy. *Regional Science and Urban Economics*, 49:217–231.

McGuire, A. (1987). The measurement of hospital efficiency. *Social Science & Medicine*, 24(9):719–724.

McKay, N. L. and Deily, M. E. (2008). Cost inefficiency and hospital health outcomes. *Health economics*, 17(7):833–848.

McKibbin, R. (2023). The effect of RCTs on drug demand: Evidence from off-label cancer drugs. *Journal of Health Economics*, 90:102779.

MoH (2014). Analisi spazio-temporale degli indicatori della Griglia LEA relativi all'adempimento sul "mantenimento dell'erogazione dei LEA" - Trend regionali anni 2001 - 2012. Technical report, Ministry of Health.

MoH (2015). Monitoraggio dei LEA attraverso la cd. Griglia LEA. Technical report, Ministry of Health.

MoH (2016). Monitoraggio dei LEA attraverso la cd. Griglia LEA. Technical report, Ministry of Health.

MoH (2017). Monitoraggio dei LEA attraverso la cd. Griglia LEA. Technical report, Ministry of Health.

MoH (2018). Monitoraggio dei LEA attraverso la cd. Griglia LEA. Technical report, Ministry of Health.

MoH (2020). Rapporto annuale sull'attività di ricovero ospedaliero - Dati SDO 2020. Technical report, Ministry of Health.

Moja, L., Piatti, A., Pecoraro, V., Ricci, C., Virgili, G., Salanti, G., Germagnoli, L., Liberati, A., and Banfi, G. (2012). Timing matters in hip fracture surgery: patients operated within 48 hours have better outcomes. a meta-analysis and meta-regression of over 190,000 patients. *Public Library of Science San Francisco, USA*.

Navarro, G., Prat-Marin, A., Asenjo, M., Menacho, A., Trilla, A., and Salleras, L. (2001). Review of the utilisation of a university hospital in Barcelona (Spain): evolution 1992–1996. *European journal of epidemiology*, 17(7):679–684.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. *Ann. Agricultural Sciences*, pages 1–51.

Nixon, J. and Ulmann, P. (2006). The relationship between health care expenditure and health outcomes: evidence and caveats for a causal link. *The European journal of health economics*, 7:7–18.

OECD (2016). *Scoping paper on health system efficiency measurement (cooperation between the OECD and EC in promoting efficiency in health care)*. OECD.

OECD (2019). *Health at a Glance 2019*. OECD.

OECD (2021). *Health at a Glance 2021*. OECD.

Palmer, S. and Torgerson, D. J. (1999). Definitions of efficiency. *Bmj*, 318(7191):1136.

Pileggi, C., Bianco, A., Di Stasio, S., and Angelillo, I. (2004). Inappropriate hospital use by patients needing urgent medical attention in Italy. *Public health*, 118(4):284–291.

Polidano, C., Zhu, A., and Bornstein, J. C. (2017). The relation between cesarean birth and child cognitive development. *Scientific reports*, 7(1):11483.

Pross, C., Busse, R., and Geissler, A. (2017). Hospital quality variation matters–A time-trend and cross-section analysis of outcomes in German hospitals from 2006 to 2014. *Health policy*, 121(8):842–852.

Quaglio, G., Karapiperis, T., Van Woensel, L., Arnold, E., and McDaid, D. (2013). Austerity and health in Europe. *Health policy*, 113(1-2):13–19.

Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, page rdad018.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512.

Robins, J. M. (1987). Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". *Computers & Mathematics with Applications*, 14(9):923–945.

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 4(3):305–22.

Roth, J., Sant'Anna, P. H., Bilinski, A., and Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Sandall, J., Tribe, R. M., Avery, L., Mola, G., Visser, G. H., Homer, C. S., Gibbons, D., Kelly, N. M., Kennedy, H. P., Kidanto, H., et al. (2018). Short-term and long-term effects of caesarean section on the health of women and children. *The Lancet*, 392(10155):1349–1357.

Santías, F. R., Cadarso-Suárez, C., and Rodríguez-Álvarez, M. X. (2011). Estimating hospital production functions through flexible regression models. *Mathematical and Computer Modelling*, 54(7-8):1760–1764.

Schiele, F., Gale, C. P., Bonnefoy, E., Capuano, F., Claeys, M. J., Danchin, N., Fox, K. A., Huber, K., Iakobishvili, Z., Lettino, M., et al. (2017). Quality indicators for acute myocardial infarction: A position paper of the acute cardiovascular care association. *European Heart Journal: Acute Cardiovascular Care*, 6(1):34–59.

SiVeAS (2014). L'erogazione dei LEA nelle Regioni in Piano di rientro: Trend 2007 - 2012 (in Italian). Technical report.

Siverskog, J. and Henriksson, M. (2022). The health cost of reducing hospital bed capacity. *Social Science & Medicine*, 313:115399.

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

Torbica, A. and Fattore, G. (2005). The "Essential Levels of Care" in Italy: when being explicit serves the devolution of powers. *The European Journal of Health Economics*, 6:46–52.

Webb, M. D. (2013). Reworking wild bootstrap based inference for clustered errors. Technical report, Queen's Economics Department Working Paper.

WHO (2006). Quality of care: a process for making strategic choices in health systems.

WHO (2018). *Handbook for national quality policy and strategy: a practical approach for developing policy and strategy to improve quality of care*. World Health Organization.

Williamson, S., Landeiro, F., McConnell, T., Fulford-Smith, L., Javaid, M. K., Judge, A., and Leal, J. (2017). Costs of fragility hip fractures globally: a systematic review and meta-regression analysis. *Osteoporosis international*, 28:2791–2800.

Wooldridge, J. M. (2021). Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators. *Available at SSRN 3906345*.

Appendix A – Additional Tables and Descriptive Statistics

Table 1: Examples of DRG at High-risk of Inappropriateness

DR	Description DRG
13	Multiple sclerosis and cerebellar ataxia
19	Cranial and peripheral nerve disorders (without complications)
36	Retina surgery
38	Iris primary surgery
39	Crystalline lens surgery with or without vitrectomy
40	Extraocular structures surgeries except for eye socket, aged > 17.
41	Extraocular structures surgeries except for eye socket, aged < 18.
59	Tonsillectomy and/or adenoidectomy (aged 17+)
60	Tonsillectomy and/or adenoidectomy (aged < 18)
133	Atherosclerosis without complications (except emergencies)
490	H.I.V. associated with or without comorbidities
563	Convulsions (aged 17+) without complications
564	Cephalalgia (aged 17+)

Table 2: Descriptive Statistics Socio-Economic Characteristics – pre- vs post-treatment period

	Never-treated			**d2007**			**d2010**			**d2011**		
	Pre 2007	**Post 2007**	**Diff-in-Means**	**Pre 2007**	**Post 2007**	**Diff-in-Means**	**Pre 2007**	**Post 2007**	**Diff-in-Means**	**Pre 2007**	**Post 2007**	**Diff-in-Means**
Population size	2401136.36	2572743.80	171607.44	2956629.41	3048436.13	91806.72	3128209.67	3172304.17	44094.50	4037581.56	4094854.08	57272.53*
	(2614284.40)	(2828880.41)	(360084.25)	(2091093.77)	(2193615.63)	(355939.07)	(1151858.00)	(1241479.77)	(371388.60)	(11962.39)	(84077.81)	(24596.54)
% people aged 15-34	25.31	20.96	-4.36***	26.73	22.40	-4.33***	26.50	22.35	-4.14***	29.14	24.15	-5.00***
	(2.01)	(1.71)	(0.25)	(2.93)	(2.63)	(0.47)	(2.86)	(2.94)	(0.90)	(1.49)	(1.46)	(0.65)
% people aged 35-54	28.89	30.53	1.64***	28.13	29.99	1.86***	28.11	29.52	1.41***	27.27	29.32	2.05***
	(1.15)	(0.85)	(0.14)	(1.24)	(1.09)	(0.20)	(1.27)	(0.89)	(0.35)	(0.71)	(0.42)	(0.27)
% people aged 55-64	12.32	12.65	0.34**	11.57	12.84	1.27***	11.80	12.87	1.07*	10.91	12.47	1.56***
	(0.91)	(0.71)	(0.11)	(1.25)	(0.81)	(0.18)	(1.57)	(0.68)	(0.39)	(0.48)	(0.45)	(0.21)
% people aged ≥ 65	20.03	22.14	2.11***	19.03	21.45	2.42***	19.29	21.83	2.54**	16.12	19.68	3.56***
	(2.31)	(2.04)	(0.29)	(3.61)	(3.25)	(0.58)	(2.32)	(2.39)	(0.73)	(0.94)	(1.47)	(0.53)
% female	51.38	51.38	-0.01	51.64	51.57	-0.07	51.39	51.41	0.03	51.44	51.48	0.03
	(0.39)	(0.38)	(0.05)	(0.57)	(0.50)	(0.09)	(0.31)	(0.24)	(0.09)	(0.04)	(0.05)	(0.02)
GDP per capita	26615.16	31334.81	4719.65***	19986.60	23425.17	3438.57***	19720.65	22926.08	3205.43	15052.73	17315.51	2262.78***
	(5194.18)	(6072.57)	(742.89)	(5455.01)	(5801.51)	(934.36)	(6166.94)	(6365.05)	(1949.59)	(1358.62)	(739.14)	(500.62)

Notes: This table presents descriptive statistics, split into pre- and post-treatment periods based on the year the first PdR was signed (2007). Means are reported with standard deviations in parentheses. Additionally, the table includes a differences-in-means test comparing averages from the pre-2007 and post-2007 periods, with standard errors (s.e.) in parentheses. **Never-treated** = Regions that have never undergone an RP. **d2007** = Regions that first signed an RP in 2007. **d2010** = Regions that first signed an RP in 2010. **d2011** = Regions that first signed an RP in 2011. ***, **, and * denote significance at 1%, 5%, and 10%, respectively.

Appendix B – Causal Inference Review

In this appendix, we provide a brief overview of how to derive the counterfactual outcome–that is, the outcome that would have been observed had the policy not been implemented–and, consequently, the average treatment effect on the treated. In particular, Subsection 1.8 reviews the main assumptions made in the program evaluation literature and their implications. Next, Subsection 1.8 discusses why the classic Two-Way Fixed Effects estimator–commonly used in the literature to estimate the impact of Recovery Plans on health-related outcomes–may be biased in settings with multiple time periods with variation in treatment timing, such as the one under analysis. Finally, Subsection 1.8 introduces the estimator proposed by Wooldridge (2021), which addresses the issues discussed in Subsection 1.8 and provides a consistent estimator of the ATT in a staggered treatment adoption framework.

Deriving the Counterfactual Outcome

To define the parameters of interest, I will use the potential outcomes framework first introduced by Neyman (1923) and later formalized by Rubin (1974). Specifically, following Callaway and Sant'Anna (2021), I will integrate the dynamic potential outcomes framework (Robins, 1986, 1987) with the dynamic treatment adoption setting (Heckman et al., 2016). For simplicity, I will adopt the same notation introduced in Section 1.6.

Let the random variable $w_{g,t}$ denote a binary treatment with support in $\{0, 1\}$. $w_{g,t} = 1$ will then denote whether unit g at time t was treated, with $g = 1, \ldots, G$ and, $t = 1, \ldots, T$.[35]

As stated by Wooldridge (2021), a staggered entry can be viewed as potentially leading to different levels of exposure to the policy, depending on when the unit was first treated. The intuition behind this latter statement is that units treated earlier are exposed to the policy for an extended period. Let q be the first period in which the policy is implemented; then, assuming there is a never treated group, one can define $T - q + 2$ mutually-exclusive cohort dummies, $d_{g,q}, \ldots, d_{g,T}$, indicating when unit g first received the treatment. Since the adoption with variation in treatment timing can be perceived as generating different intensities of the treatment effect, a way to model this problem is to exploit an expanded set of potential outcomes (Wooldridge, 2021). For $r \in \{q, \ldots, T\}$, $y_{g,t}(r)$ will represent the potential outcome for unit g at time t had the policy been introduced by period r (i.e., $d_{g,r} = 1$), whereas $y_{g,t}(\infty)$ will denote the analogous in period t had the treatment not been received (that is, had the unit never been treated).

Using the above notation, the causal effect of the policy at time t for a specific unit g first receiving the treatment in r will be given by $te_{g,t} = y_{g,t}(r) - y_{g,t}(\infty)$.[36] Then, the realized outcome for a generic unit can be rewritten as:

$$\begin{aligned} y_{g,t} =& y_{g,t}(\infty) + d_{g,q} \cdot [y_{g,t}(q) - y_{g,t}(\infty)] + d_{g,q+1} \cdot [y_{g,t}(q+1) - y_{g,t}(\infty)] \\ &+ \ldots + d_{g,T} \cdot [y_{g,t}(T) - y_{g,t}(\infty)] \end{aligned}$$

[35] For simplicity, to keep notation in line with that introduced in Section 1.6, units will be denoted with g and. In addition, without any loss of generality, g will also represent the level at which standard errors will be clustered in the following subsections. Indeed, it is common in the DiD literature to cluster standard errors at the level at which the treatment is assigned (Cameron and Miller, 2015).

[36] As pointed out by Wooldridge (2021), other possible treatment effects can be retrieved in this framework. For instance, another possibility would be to estimate the treatment effect given by $te_{g,t} = y_{g,t}(r) - y_{g,t}(r+1)$ for $t \geq r$. This latter represents the cumulative effect of being first treated in period r rather than in period $r + 1$.

where $d_{g,q} + d_{g,q+1} + \ldots + d_{g,T} + d_{g,\text{never}} = 1$. Oftentimes, however, the causal estimand of interest is the average treatment effect on the treated, ATT, in periods where the treated cohorts are effectively under the policy. That is:

$$\tau_{r,t} = \mathbb{E}(te_{g,t}(r)|d_r = 1), \quad r \in \{q, \ldots, T\}; \quad t = r, \ldots, T \tag{B.1}$$

where to have at least a never treated period, it is usually imposed that $q \geq 2$.

The problem of how to estimate $y_{g,t}(\infty)$ represents the fundamental problem of causal inference (Holland, 1986). If, for a moment, we abstract from the fact that we are in a staggered intervention setting and think in terms of a common entry setup,[37] then units can be either treated or not, and the complete set of potential outcomes shrinks to two mutually-exclusive potential outcomes (i.e., either $y_{g,t}(r)$ or $y_{g,t}(\infty)$). Then, as noticed by Manski and Pepper (2013), the fact that data will reveal only one of the two mutually-exclusive quantities constitutes the selection problem. While $y_{g,t}(r)$ will be observed for all treated units, the problem will be to find a way to infer the benchmark outcome unit g would have experienced at time t had the policy not been introduced by period r. That is, $y_{g,t}(\infty)$.

In the program evaluation literature, $y_{g,t}(\infty)$ is usually retrieved by invoking an "exact" invariance assumption (Imbens and Wooldridge, 2009; Depalo, 2019). Following Depalo (2019), these assumptions can be divided into four main groups. The *time invariance* assumption exploits the outcome observed in the pre-treatment period for the treated unit g to estimate the benchmark outcome. That is, $\hat{y}_{g,t}(\infty) = y_{g,pre}(\infty)$ where $pre \in \{1, \ldots, q\}$ and $q = r - 1$. Another common way to retrieve the counterfactual outcome, $y_{g,t}(\infty)$, is to rely on a *state invariance* assumption, which uses (often a linear combination of) the observed outcomes in the never-treated units. If $never$ denotes the set of never-treated units, then for $t \geq r$, one way to retrieve $\hat{y}_{g,t}(\infty)$ is either to impose $\hat{y}_{g,t}(\infty) = y_{never,t}(\infty)$ or $\hat{y}_{g,t}(\infty) = \mathbb{E}(y_{never,t}(\infty))$. Depending on the setting at hand (common entry vs. staggered rollout) and whether it is more plausible that the assumption is valid only after conditioning on observable characteristics or not, one can invoke a "suitable" *parallel trends* assumption to estimate the benchmark.[38] This assumption requires that had the policy not been implemented, the average outcome evolution for the treated and not-yet-treated – or never-treated, depending on which reference group is considered – would have remained the same over time (e.g., in the case of an unconditional parallel trends based on a never-treated group this would imply that $\mathbb{E}(y_t(\infty) - y_1(\infty)|d_q, \ldots, d_T) = \mathbb{E}(y_t(\infty) - y_1(\infty))$).[39] The last group includes the *time-varying parallel trends* assumption. To retrieve the counterfactual, $y_{g,t}(\infty)$, a weighted average of the units in the donor pool period is used, where the weights are chosen appropriately according to the Synthetic Control Method first proposed by Abadie and Gardeazabal (2003) (i.e., $\hat{y}_{g,t}(\infty) = \mathbf{w}y_{never,t\geq r}(\infty)$ where $\mathbf{w}$ is the vector of selected weights). For a review of the Synthetic Control Method, please refer either to Appendix A of Depalo (2019), in which the author briefly reviews how this procedure works and its potential pitfalls, or to Abadie (2021).

[37] Please note that this simplification is just for expositional convenience and does not affect the following analysis.

[38] Given the recently developed Difference-in-Differences (DiD) estimators, depending on the problem at hand (e.g., common entry vs. variation in treatment timing) and the estimator employed to carry out the policy evaluation exercise, there exist different types of common trends assumptions. Some variations of this assumption will be explained in the Section 1.8.

[39] Please note that if a common entry setup is considered, then there are only two mutually exclusive groups: treated and untreated. Thus, the distinction between never-treated and not-yet-treated units is redundant.

TWFE & DiD Decomposition

While it is common practice to estimate a Difference-in-Differences model using a Two-Way Fixed Effects regression, as presented below, recent advancements in the program evaluation literature have shown that the estimated coefficient of the treatment dummy in a TWFE model is an inconsistent estimator of the ATT in settings with variation in treatment timing (see, for instance, Borusyak et al., 2021; Goodman-Bacon, 2021; De Chaisemartin and d'Haultfoeuille, 2022a).

Using the notation introduced above, let us consider the following regression model:

$$y_{g,t} = \beta w_{g,t} + c_g + \eta_t + u_{g,t} \quad g = 1, \ldots, G, \quad t = 1, \ldots, T \tag{B.2}$$

where c_g are unit-specific fixed effects (FE), η_t are year FE, and $u_{g,t}$ is the error term. It can be shown that the estimator for β, in the context of a multi-period setup with variation in treatment timing, does not identify the ATT anymore. The equivalence between the canonical (2×2) Difference-in-Differences (DiD) and the TWFE no longer holds (Imai and Kim, 2021). In the context of a staggered setup with treatment at an absorbing state – that is, once the unit g receives the treatment, it remains treated for the remainder of the panel (i.e., $w_{g,s} \leq w_{g,t}$ for $s < t$) – Goodman-Bacon (2021) shows that the estimator for β in (B.2) is a convex-weighted average of all the possible combinations of (2×2) DiD estimators.

To give the intuition behind this result, for simplicity, suppose that the panel is balanced with T periods and G cross-sectional units, only three groups exist, and there are no leavers: units can be either never-treated or receiving the treatment in period k (early treated) – where $k \in t = 2, \ldots, T$ (i.e., there is at least one period in which all units are untreated) – or in period $l > k$ (later treated). Goodman-Bacon (2021) proves that there will exist overall four pairs of (2×2) DiD estimators in this case.[40] As shown in Figure .6, if Eq. (B.2) is estimated using only either units treated in period k and never-treated units (Panel A) or units treated in period l and never treated group (Panel B), then the TWFE reduces to the classic DiD estimator. In particular, following Goodman-Bacon (2021), it can be proved that:

$$\hat{\beta}^{2\times2}_{j,never} \equiv \left(\bar{y}_j^{POST(j)} - \bar{y}_j^{PRE(j)}\right) - \left(\bar{y}_{never}^{POST(j)} - \bar{y}_{never}^{PRE(j)}\right), \quad j = k, l \tag{B.3}$$

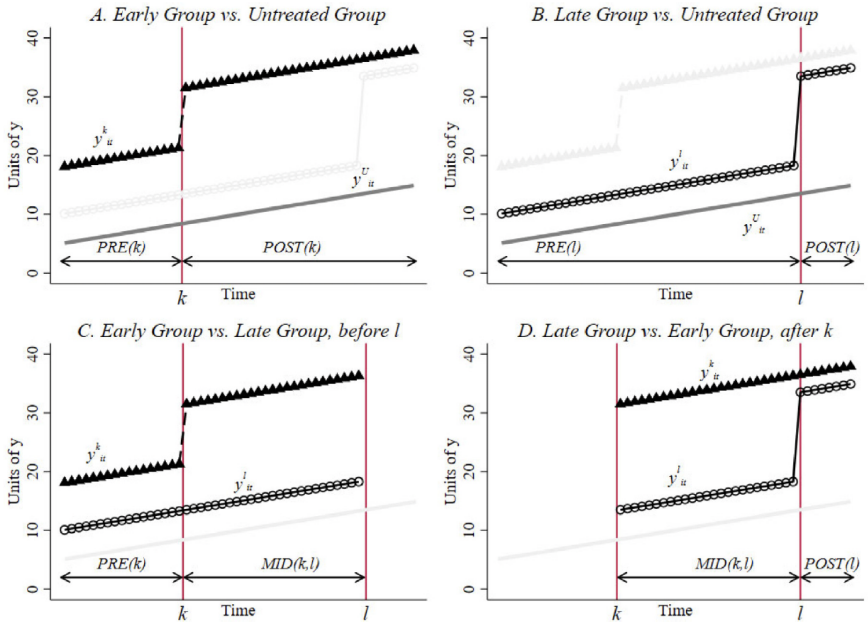
Panels C and D of Figure .6, instead, plot what would happen if Eq. (B.2) is estimated using only timing groups (that is, groups k and l). Identification in these two plots comes from these two groups receiving the treatment in different periods. Specifically, before l, the early-treated group act as the treatment group since it is the group that experiences a switch in the treatment status at period $k < l$ (Panel C). At the same time, the later-treated group serves as a benchmark for the early-treated since it is still not under treatment in this period. It can be shown that the estimated coefficient for β obtained by estimating Eq. (B.2) using these two groups is equal to:

$$\hat{\beta}^{2\times2,k}_{k,l} \equiv \left(\bar{y}_k^{MID(k,l)} - \bar{y}_k^{PRE(k)}\right) - \left(\bar{y}_l^{MID(k,l)} - \bar{y}_l^{PRE(k)}\right) \tag{B.4}$$

where now the time window considered goes from the pre-treatment period in which group k is

[40] The author uses the term (2×2) to refer to a pair made up of a group whose treatment status changes over the observed period and a group whose treatment status is, instead, stable and a time window of two periods (PRE and $POST$ the introduction of the policy)

Figure .6: Source: p. 257 Goodman-Bacon (2021)



still not treated and the period $MID(k,l)$, where group k is treated, but group l is not.

In Panel D, early-treated units act as a control for those receiving the treatment in period l. This latter pair is often referred to in the DiD literature as a "forbidden comparison", as later-treated will be compared to already-treated units. In this latter case, the estimator for β will be:

$$\hat{\beta}^{2\times2,l}_{k,l} \equiv \left(\bar{y}^{POST(l)}_{l} - \bar{y}^{MID(k,l)}_{l}\right) - \left(\bar{y}^{POST(l)}_{k} - \bar{y}^{MID(k,l)}_{k}\right) \tag{B.5}$$

Please note that each subsample employs a fraction of the full sample. To estimate Eq. (B.3) only two groups out of four are employed. By contrast, all time periods are used. This implies that their sample shares will be equal to $n_k + n_{never}$ and $n_l + n_{never}$, where $n_j = \sum_g \mathbb{1}\{t_g = j\}/G, \quad j = k, l, never$. To estimate Eq. (B.4), also two groups are considered, but only some periods. In this case, the subsample share will be equal to $(n_k + n_l)(1 - \bar{w}_l)$, where $\bar{w}_l \equiv \sum_t \mathbb{1}\{t \geq l\}/T$, which is the share of time for which groups l remains treated. Lastly, only two groups and some periods are used to estimate Eq. (B.5). Overall, the subsample share used to obtain $\hat{\beta}^{2\times2,l}_{k,l}$ amounts to $(n_k + n_l)\,\bar{w}_k$.

The author also shows that it is possible to quantify the amount of identifying variation used to estimate (B.3)– (B.5). This "equals the variance of fixed-effects-adjusted $w_{g,t}$ from its subsamples" (Goodman-Bacon, 2021, p. 257):

$$\hat{V}^{w}_{j,never} \equiv n_{j,never}\,(1 - n_{j,never})\,\bar{w}_j\,(1 - \bar{w}_j), \quad j = k, l \tag{B.6}$$

$$\hat{V}^{w,k}_{k,l} \equiv n_{k,l}\,(1 - n_{k,l})\,\frac{\bar{w}_k - \bar{w}_l}{1 - \bar{w}_l}\,\frac{1 - \bar{w}_k}{1 - \bar{w}_l} \tag{B.7}$$

$$\hat{V}^{w,k}_{k,l} \equiv n_{k,l}\,(1 - n_{k,l})\,\frac{\bar{w}_l}{\bar{w}_k}\,\frac{\bar{w}_k - \bar{w}_l}{\bar{w}_k} \tag{B.8}$$

where $n_{a,b} \equiv \frac{n_a}{n_a + n_b}$ represents the relative size, in each (2×2) pair, of the group that receives the

treatment.

In each of the pairwise variance formulas above, if either $n_{j,never}$ or $n_{k,l}$ goes to 0 or 1, then the variance degenerates to 0. This is because there will not be either a treatment or a control group. The third and the fourth terms in the variance formulas, instead, tell the time in which the treatment is assigned in that specific subsample. Since $\bar{w}$ represents the variance of $w_{g,t}$ in each (2×2) pair, if this goes to 0 or 1, then the treatment status does not change over the observed period.

Goodman-Bacon proves that the TWFE estimator for β is a weighted average of these four (2×2) DiD estimators, such as those in (B.3)– (B.5), where weights are proportional to subsample shares and the variance of $w_{g,t}$ (where the variance is highest for units treated in the middle of the panel).

This reasoning can be generalized to more than three groups. Using the notation introduced in the previous subsection, suppose there exists a never treated group and that there are new treated units in each period r with $r \in \{q, \dots, T\}$ and $q > 1$. Then the author proves that:

$$\hat{\beta} = \sum_r s_{r,never} \hat{\beta}^{2\times2}_{r,never} + \sum_r \sum_{l>r} [s^r_{r,l} \hat{\beta}^{2\times2,r}_{r,l} + s^l_{r,l} \hat{\beta}^{2\times2,l}_{r,l}] \tag{B.9}$$

where, as before:

$$\hat{\beta}^{2\times2}_{r,never} \equiv \left(\bar{y}^{POST(r)}_r - \bar{y}^{PRE(r)}_r\right) - \left(\bar{y}^{POST(r)}_{never} - \bar{y}^{PRE(r)}_{never}\right)$$

$$\hat{\beta}^{2\times2,r}_{r,l} \equiv \left(\bar{y}^{MID(r,l)}_r - \bar{y}^{PRE(r)}_r\right) - \left(\bar{y}^{MID(r,l)}_l - \bar{y}^{PRE(r)}_l\right)$$

$$\hat{\beta}^{2\times2,l}_{r,l} \equiv \left(\bar{y}^{POST(l)}_l - \bar{y}^{MID(r,l)}_l\right) - \left(\bar{y}^{POST(l)}_r - \bar{y}^{MID(k,l)}_r\right)$$

and the weights are equal to:

$$s_{r,never} = \frac{(n_r + n_{never})^2 \hat{V}^w_{r,never}}{\hat{V}^w}$$

$$s^r_{r,l} = \frac{((n_r + n_{never})(1 - \bar{w}_l))^2 \hat{V}^{w,r}_{r,l}}{\hat{V}^w}$$

$$s^l_{r,l} = \frac{((n_r + n_{never})\bar{w}_r)^2 \hat{V}^{w,r}_{r,l}}{\hat{V}^w}$$

and $\sum_r s_{r,never} + \sum_r \sum_{l>r} [s^r_{r,l} + s^l_{r,l}] = 1$ where, again, $r \in \{q, \dots, T\}$.

Overall, there will be at most $T-q+2$ possible pair of (2×2) DiD estimators. Two points are worth to be noted. First, weights depend on the size of the subsamples (squared) and the subsample variances in (B.6)–(B.8), where the variances will be larger whenever either the two groups are approximately similar in size or when the treatment occurs in the middle of the time window. Second, by simply modifying the dimension of the panel under analysis, the estimate of β can change dramatically even if the 2×2 DiD estimators are constant.

In the limit, the author proves that:

$$\operatorname*{plim}_{G\to\infty} \hat{\beta} = VWATT + VWPT - \Delta ATT$$

Where $VWATT$ represents a variance-weighted average of treatment effects, the second term, $VWPT$, comes from the fact that each pair of DiD relies on a pairwise parallel trends assumption for identification. Thus $VWPT$ represents a generalization of the parallel trends assumption to the staggered entry setup. The last term, ΔATT, is the variation in the treatment effect. It is immediate to see from the Goodman-Bacon's *DiD decomposition* that, even if there is no treatment effect heterogeneity, the estimated coefficient from the classic TWFE regression is not estimating the ATT, but a variance-weighted version of it.

In the absence of treatment effect heterogeneity (across either time or units), the $\hat{\beta}$ can be proved to be a variance-weighted average of ATT, with all weights being positive. Conversely, if the treatment effect is likely to be heterogeneous, then the problem of negative weights arises. This is because when already-treated units serve as controls for the later-treated, "changes in their outcomes are subtracted, and these changes may include time-varying treatment effects" (Goodman-Bacon, 2021, p. 2).

Two-Way Mundlak Approach (TWM)

Different estimators have been recently proposed to prevent the issue of negative weights discussed in Section 1.8 and thus to retrieve a consistent estimator of the ATT in the context with staggered treatment adoption. For instance, see Borusyak et al. (2021); Callaway and Sant'Anna (2021); Sun and Abraham (2021); Wooldridge (2021); De Chaisemartin and d'Haultfoeuille (2022a).[41]

However, Wooldridge (2021) shows there is nothing intrinsically wrong with the TWFE estimator. The main problem with this estimator is that it is often applied to a model too restricted in the number of parameters. He proves the equivalence between the TWFE estimator and the Pooled OLS (POLS) estimator applied to a regression which includes time-specific cross-sectional averages and unit-specific time averages, the *Two-way Mundlak* approach (TWM).[42] This result is paramount as it allows us to understand better how the TWFE works. Besides, when applied to carry out policy evaluation exercises, the TWM tool permits retrieving a consistent estimator of the ATT within the staggered treatment adoption setting.

Depending on whether there exist never-treated units or not, Wooldridge shows that different ATT can be retrieved. To save space, and since in the context described in Section 1.3, there exists more than one untreated unit, how to infer something about the average treatment effects on the treated presented in (B.1) will only be explained.

Wooldridge shows that a consistent estimator for (B.1) can be retrieved by estimating the following regression via POLS:

$$y_{g,t} = \alpha + \sum_{r=q}^{T}\sum_{s=r}^{T} \tau_{r,s}(w_{g,t} \cdot d_{g,r} \cdot fs_t) + \sum_{r=q}^{T} \lambda_r d_{g,r} + \eta_t + u_{g,t} \tag{B.10}$$

where α is the constant, η_t are year FE, and fs_t is a dummy variable equal to 1 if $s = t$ and zero otherwise. It should be noticed that $w_{g,t} \cdot d_{g,r} \cdot fs_t = d_{g,r} \cdot fs_t$ for $s \geq r$. Including $w_{g,t}$ in (B.10) highlights that it is still possible to obtain a consistent estimator of the $ATTs$ in a staggered rollout context. Further, it shows that considerable heterogeneity can be allowed within the simple

[41] Please refer to Roth et al. (2023) and De Chaisemartin and d'Haultfoeuille (2022b) for an in-depth review of all the recently developed DiD-type estimators.

[42] This equivalence holds only in the context of balanced panels. However, this does not represent a threat in the context of RPs, as the panel under analysis is balanced.

linear regression framework. An equivalent TWFE estimator can be obtained by applying the within estimator to (B.10), after having dropped $d_{g,q}, \ldots, d_{g,T}$. This is because, in (B.10), d_r represents the time average of $w_{g,t}$ (for all treated units in a given cohort). While, for each t, fs_t's represent the cross-sectional averages.

To estimate consistently $\tau_{r,t}$, besides requiring that the treatment is at an absorbing state (i.e., no leavers) and that the Stable Unit Treatment Value Assumption hold, two additional assumptions are needed for identification. The first rules out anticipatory behaviors.

No Anticipation (NA): For each treatment cohort $r \in \{q, \ldots, T\}$

$$\mathbb{E}(y_t(r) - y_t(\infty)|d_q, \ldots, d_T) = 0, \quad \forall t < r$$

This means that, on average, the potential outcomes between treated and never-treated units are the same in the pre-intervention period, regardless of when a unit is first treated. This is similar to the strict exogeneity assumption required to estimate FE in panel data models.

The second assumption is a generalization of the parallel trends assumption to the multi-period setup with variation in treatment timing.

Parallel Trends (PT): For each d_r with $r \in \{q, \ldots, T\}$

$$\mathbb{E}(y_t(\infty) - y_1(\infty)|d_q, \ldots, d_T) = \mathbb{E}(y_t(\infty) - y_1(\infty))$$

where $t = \{2, \ldots, T\}$. This assumption requires the average evolution in the benchmark state to be mean independent of the treatment status. The common trends (CT) assumption can be stated equivalently in terms of adjacent periods as follows:

$$\mathbb{E}(y_t(\infty) - y_{t-1}(\infty)|d_q, \ldots, d_T) = \mathbb{E}(y_t(\infty) - y_{t-1}(\infty))$$

where in each of the two versions of the PT assumption, it is implicitly assumed that for each $r \in q, \ldots, T$, there is a positive probability that some units are receiving the treatment. If, for instance, for $r = q + 3$, there is no unit entering the treatment, then $\tau_{q+3,t}$ cannot be identified.

Suppose the researcher believes the PT or NA assumption is unlikely to hold. One of the main advantages of the TWM approach is that it allows the researcher to relax these assumptions by conditioning on observable characteristics and allowing the ATT to vary with them.

Suppose $\mathbf{x}_g$ denotes a vector of (time-invariant) covariates. The NA assumption can be modified such that the TE should be zero for each subpopulation defined by $\mathbf{x}$ (Wooldridge, 2021).

Conditional No Anticipation (CNA): For each treatment cohort $r \in \{q, \ldots, T\}$,

Similarly, the PT assumption can be modified as follows:

$$\mathbb{E}(y_t(r) - y_t(\infty)|d_q, \ldots, d_T, \mathbf{x}) = 0, \quad \forall t < r$$

Conditional parallel trends (CPT): For each d_r with $r \in \{q, \ldots, T\}$ and covariates $\mathbf{x}$,

$$\mathbb{E}(y_t(\infty) - y_1(\infty)|d_q, \ldots, d_T, \mathbf{x}) = \mathbb{E}(y_t(\infty) - y_1(\infty)|\mathbf{x})$$

Accordingly, the causal estimands of interest become:

$$\tau_{r,t}(\mathbf{x}) = \mathbb{E}(te_t(r)|d_r = 1, \mathbf{x})$$

Which are the ATT_s after having conditioned on observable characteristics. Assuming also that the model is linear in the parameter, then it is possible to estimate the coefficients of the following regression consistently via POLS:

$$\begin{aligned} y_{g,t} =& \alpha + \sum_{r=q}^{T}\sum_{s=r}^{T}\tau_{r,s}(w_{g,t}\cdot d_{g,r}\cdot fs_t) + \mathbf{x}_g\boldsymbol{\kappa} + \sum_{r=q}^{T}(d_{g,r}\cdot\mathbf{x}_g)\boldsymbol{\gamma}_r + \sum_{s}^{T}(fs_t\cdot\mathbf{x}_g)\boldsymbol{\pi}_s \\ & \sum_{r=q}^{T}\sum_{s=r}^{T}(w_{g,t}\cdot d_{g,r}\cdot fs_t\cdot\dot{\mathbf{x}}_{g,r})\boldsymbol{\rho}_{r,s} + \sum_{r=q}^{T}\lambda_r d_{g,r} + \eta_t + u_{g,t} \end{aligned} \tag{B.11}$$

where $s = \{2,\dots,T\}$, $\dot{\mathbf{x}}_{g,r} = (\mathbf{x}_g - \boldsymbol{\mu}_r)$, and $\boldsymbol{\mu}_r = \mathbb{E}(\mathbf{x}_g|d_r = 1)$. The idea to center $\mathbf{x}$ about the mean of the x's over the treated cohorts ensures that $\tau_{g,t}$ represents the ATT (Wooldridge, 2021).[43] If all the terms involving d_r, $\mathbf{x}_g$, and $d_{g,r}\cdot\mathbf{x}_g$ are dropped, an equivalent TWFE estimator to that obtained by applying POLS to (B.11) can still be retrieved.

It is easy to understand how much flexibility can be introduced within the simple linear regression framework from eq. (B.11). If, on the one side, this guarantees that a consistent estimator of the ATT's can still be retrieved within the linear regression framework, on the other hand, even if the dimension of T is moderate, the number of parameters to be estimated in (B.11) is huge. Specifically, it could be the case that for some $r \in q,\dots,T$, no new units are being treated (that is, some cohorts may not exist), causing some of the $\tau_{r,t}$ not to be identified. Even if there are new units receiving the policy for each $r \in q,\dots,T$, it could be that only a tiny fraction of them enters the treatment in a particular period, causing the $\tau_{r,t}$'s – and consequently, confidence intervals – to be imprecisely estimated.

To overcome this issue, Wooldridge proposes two solutions. Either to estimate (B.11) and then aggregate them in a small number of TE – taking, for instance, a linear combination – or to impose ex-ante restrictions on the number of parameters to be estimated. For instance, one could group treated units into two mutually-exclusive cohorts (early vs. later treated units), or one can allow the ATT to vary only over time by imposing homogeneity across treated cohorts. Another possibility could be to require homogeneity over time, allowing the ATT to vary by cohorts or imposing restrictions on the covariates, for example, by requiring them to vary only across treatment status.

As pointed out by the author, ex-ante restrictions can be easily tested by estimating the unrestricted and the restricted models and then constructing a Wald statistic to test the exclusion restrictions.

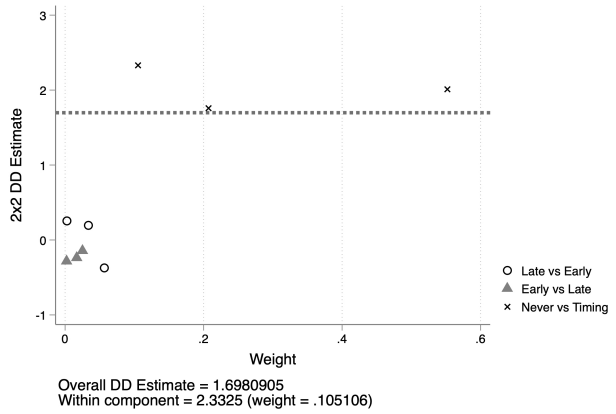
[43] Note it is sufficient to de-mean the x's only when interacted with $w_{g,t}$.

Appendix C – DiD Decomposition

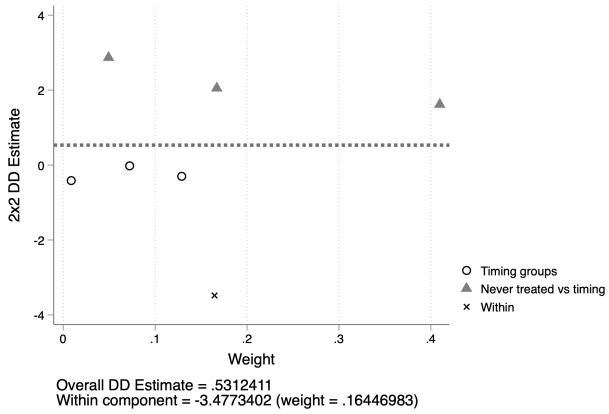
Most existing studies on RPs rely on variations of the classic TWFE estimator to evaluate the impact of the policy on various health indicators. However, as highlighted in Appendix B, failing to account for the staggered nature of the policy can invalidate inference in the presence of heterogeneous treatment effects. Furthermore, Goodman-Bacon (2021) demonstrates that, even if treatment effects are homogeneous and the Parallel Trends assumption is constant over time, the coefficient for the treatment dummy in a TWFE regression does not provide a consistent estimator for the ATT but rather for a variance-weighted version of it.

This subsection employs the DiD decomposition to illustrate the importance of considering the staggered nature of PdRs. To save space, only results for the mortality rate from ischaemic heart diseases will be shown. The intuition behind the results obtained for this indicator is similar to that for the other variables considered (available upon request).

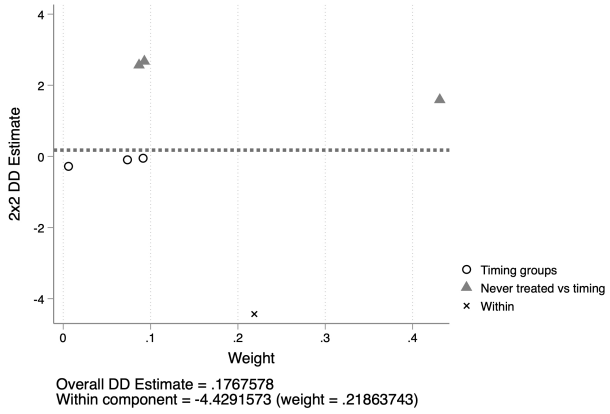
Figure C.1: DiD Decomposition – Mortality rate from ischaemic heart diseases



(a) Without controls



(b) With controls



(c) Restricted sample + controls

Panel (a) of Figure C.1 plots each pair of (2×2) DiD estimators against their weights, which are derived from applying the DiD decomposition to the estimated coefficient for β in (B.2), using the mortality rate from ischaemic heart diseases as dependent variable.[44] Specifically, the crosses represent

[44] These graphs were obtained using the STATA package `bacondecomp` provided by Goodman-Bacon.

terms where never-treated regions act as controls, and regions treated at some point act as treated. The open circles represent terms where later treated units act as controls for early treated regions (for instance, Calabria, Piemonte, and Puglia act as a control for Abruzzo in 2007). The closed triangles represent terms where early treated units (e.g., those treated in 2007) serve as controls for later treated regions (those joining an RP in 2010 or 2011), known as *forbidden comparison* groups. The dashed line is the overall TWFE estimate for β, approximately equal to 1.70 p.p.

Plotting each pair of (2×2) DiD estimators against their weights allows the researcher to assess the bias visually by inspecting which of the three types of groups (*Early vs. Late*, *Late vs. Early*, *Never vs. Timing*) receive more weights. Summing the weights of each comparison group shows how much of the identifying variation comes from a specific group, helping to evaluate how well the TWFE estimator performs relative to other estimators.

Overall, what emerges is that most of the identifying variation – equal to 86% of the overall variation – comes *Never vs. Timing* terms, for an average effect of 1.99p.p.. Despite only 4% of the identifying variation coming from terms in which early treated regions act as control groups for later treated units, the estimated impact for these terms is of the opposite sign compared to the DiD terms obtained using as controls never-treated regions. This suggests that the estimated coefficient for β in (B.2) is (slightly) biased downward.

Panel (b) presents the DiD decomposition applied to a variation of (B.2) that controls for the variables presented in Section 1.5. In this graph, the open circles represent timing regions (i.e., Early vs. Late and Late vs. Early), the closed triangles are the terms in which never-treated regions act as the control group for treated ones, and the crosses represent the *within* variation. This latter source of variation arises from the inclusion of time-varying controls in (B.2), which accounts for differences in covariates across regions within the same timing group.

In this case, almost 63% of the overall identifying variation comes from *Never vs. Timing* terms, with an average effect of $1.83 p.p.$. 16% of the variation now comes from the covariates, with an (overall) average effect of -3.48. The remaining 20% of the identifying variation comes from the timing terms, for an average effect of $-.20$. As in previous panels, the bias results in a downward bias in the estimated coefficient.

Finally, since the DiD decomposition assumes that treated units are at an absorbing state, Panel (c) reports the same results shown in Panel (b), but excludes regions that left the treatment status: Liguria, Piemonte, and Sardegna. Although the overall coefficient is smaller than that in Panel (b), the estimated coefficient remains biased downward due to the timing groups.

These results demonstrate that the estimated coefficient obtained using the classic TWFE estimator is biased in this context. As previously noted, ignoring the staggered nature of the policy would undermine inference.

Appendix D – Statistical Inference

Statistical Inference

In the context of clustered data with a balanced number of observations per cluster, the consistency of the cluster-robust variance-covariance estimator (CRVE) hinges upon the fact that, as the number of clusters, G, goes to infinity, the distribution of the cluster-robust statistic at hand approaches the actual distribution.[45] However, when G is small, the cluster-robust t-statistic can severely over-reject (MacKinnon and Webb, 2018). In this latter case, more reliable inference can be attained by using a bootstrap approximation.

To overcome the issue of over-rejection, Cameron et al. (2008) propose using a bootstrap procedure that maintains regressors fixed across bootstrap replications, the *wild cluster bootstrap* (WC). Monte Carlo simulation results in Cameron et al. (2008) suggest that, in a context with few clusters, the WC bootstrap solves the problem of over-rejection that not even the percentile bootstrap tends to eliminate. Using the notation introduced above, let $\mathbf{z}_{g,t} = (w_{g,t} \quad c_g \quad \eta_t)$ denote a row vector including all the regressors in (B.2), with $g = 1, \ldots, G$ and $t = 1, \ldots, T$. Let $\boldsymbol{\theta}$ be a column vector containing all the coefficients in (B.2). Then (B.2) can be rewritten as follows:

$$y_{g,t} = \mathbf{z}_{g,t}\boldsymbol{\theta} + u_{g,t} \tag{D.1}$$

The scheme proposed by Cameron and Miller (2015) will be followed to give the main intuition behind the WC bootstrap. Suppose it is of interest testing whether β in (B.2) is statistically different from 0, then the WC bootstrap works as follows. First, estimate (D.1) by imposing $H_0 : \beta = 0$ to obtain an estimate of $\tilde{\boldsymbol{\theta}}_{H_0}$. Second, derive the t^{th} residual within cluster g, $\tilde{u}_{g,t} = y_{g,t} - \mathbf{z}_{g,t}\tilde{\boldsymbol{\theta}}_{H_0}$. If the bootstrap procedure is replicated B times, for each b^{th} replication one should:

1a) Assign cluster g a weight, t_g, following the two-point Rademacher distribution taking values in the support in $\{-1, 1\}$, where $prob(t_g = -1) = prob(t_g = 1) = 0.5$. All observations for unit g will receive the same value of t_g.

1b) Create the pseudo-residuals, $u^*_{g,t}$, as $u^*_{g,t} = t_g \times \tilde{u}_{g,t}$. Then the new outcome variables, $y^*_{g,t}$, can be generated as $y^*_{g,t} = \mathbf{z}_{g,t}\tilde{\boldsymbol{\theta}}_{H_0} + u^*_{g,t}$

2) Obtain an estimate of $\hat{\beta}^*_b$ for the b^{th} resample by regressing $y^*_{g,t}$ on $\mathbf{z}_{g,t}$.

3) Compute the test $test^*_b = \frac{(\hat{\beta}^*_b - \hat{\beta})}{se(\hat{\beta}^*_b)}$, where $se(\hat{\beta}^*_b)$ is the standard error of $\hat{\beta}^*_b$, whereas $\hat{\beta}$ represents the estimate of β obtained using the full sample.

Then the bootstrapped p-value will be the fraction of times that $|test| > |test^*_b|$, where $b = 1, \ldots, B$.

Unlike the pair bootstrap, regressors are now kept fixed in each resample. By stacking observations for the g^{th} cluster in a vector, it can be shown that having regressors fixed implies that, for each draw, $\mathbf{y}^*_g$ will either be equal to $\mathbf{y}^*_g = \mathbf{Z}_g\tilde{\boldsymbol{\theta}}_{H_0} + \tilde{\mathbf{u}}_g$ or $\mathbf{y}^*_g = \mathbf{Z}_g\tilde{\boldsymbol{\theta}}_{H_0} - \tilde{\mathbf{u}}_g$. Webb (2013) shows that only $2^{(G-1)}$ possible values of $test^*_1, \ldots, test^*_B$ can be obtained at most. This causes problems when G is small,

[45] For a thorough review of the cluster-robust literature, please refer to Cameron and Miller (2015) and MacKinnon et al. (2023).

as the researcher may end up choosing "just one point from the interval of equally plausible p-values" (Cameron and Miller, 2015, p. 27). To avoid this issue, when $G < 10$, Webb (2013) proposes using a six-point distribution for t_g with equally-probable values, $\{-\sqrt{1.5}, -1, -\sqrt{.5}, \sqrt{.5}, 1, \sqrt{1.5}\}$.

Simulation-based evidence in MacKinnon and Webb (2017) suggests that the WC bootstrap also performs well when $15 \leq G \leq 20$, with rejection frequencies close to nominal levels (provided that the size of clusters is approximately equal). Conversely, this procedure fails when inference is carried out on a dummy variable taking value 1 only for a few clusters. In such a case, MacKinnon and Webb (2018) finds that the test based on the unrestricted wild cluster bootstrap (WCU) – obtained by not imposing H_0 – can lead to severe over-rejection when the number of treated cluster, G_1, is either equal to 1 or 13 but performs fairly well for $6 \leq G_1 \leq 8$. In contrast, tests based on the restricted wild cluster (WCR) bootstrap fail to reject when $G_1 = 1$ or $G_1 = 13$, under-reject when $G_1 = 2$ or $G_1 = 12$, work fairly well when $G_1 = 11$, and over-reject for other small values of G_1.

To give the intuition why the WC bootstrap fails in a context with few treated clusters, the same example provided by MacKinnon and Webb (2018) will be used. Suppose $G_1 = 1$, then each observation within the treated cluster will be assigned the same value of t_g, implying it is always the case that $\mathbf{u}_g^* \propto \tilde{\mathbf{u}}_g$. If, on one side, this peculiarity of the WC bootstrap ensures that, for each cluster, the bootstrapped disturbances mimic the variance-covariance structure of the true disturbances. Conversely, when G_1 is small, inference may be imprecise.

To overcome this issue, MacKinnon and Webb (2018) suggest relying on a variation of the WC bootstrap, the *subcluster wild bootstrap*. The main goal of this procedure is to eliminate the dependence of $\mathbf{u}_g^*$ on $\tilde{\mathbf{u}}_g$. To do so, the subcluster wild bootstrap still relies on a version of the CRVE to obtain cluster-robust standard errors. However, rather than multiplying each observation within cluster g by t_g, this bootstrap procedure partitions $\tilde{\mathbf{u}}_g$ into mutually-disjoint subclusters and multiplies each of them by a random weight. When each subvector is a singleton, the subcluster wild bootstrap will converge to the ordinary bootstrap.

MacKinnon and Webb (2018) find that, when G_1 is small, the ordinary wild bootstrap may lead to improved finite-sample inference if the following three assumptions are satisfied:

1. Cluster sizes should be equal and the sample size fixed.

2. The average within-cluster correlation should be small when the number of treated clusters is small.

3. The covariance matrices Ω_g need to be proportional, but between-cluster heteroskedasticity is allowed. That is, $\Omega_g = \lambda_g \bar{\Omega}$, where $\bar{\Omega}$ is a positive-definite matrix, λ_g is a scalar factor with $\lambda_1 = 1$, and $\lambda_g > 0$.

It is easy to check whether Assumption 1 is satisfied, as both the sample dimension and the cluster sizes are directly observable from the researcher. As far as Assumption 3 is concerned, MacKinnon and Webb (2018) state that it is likely to hold as any cross-cluster heteroskedasticity is allowed. Conversely, Assumption 2 is more difficult to check whether it holds in practice. However, Monte Carlo simulations in MacKinnon and Webb (2018) show that potential violations to Assumption 2 do not significantly affect inference when $G_1 \geq 2$, provided that the within-cluster correlation is not large. Furthermore, the authors also show that one way to check whether Assumption 2 holds is to look at the p-values of

the restricted and unrestricted ordinary wild bootstrap. If the p-values are similar, with the p-value of the restricted wild bootstrap being larger than the p-value of the unrestricted, Assumption 2 is likely to hold. For a thorough discussion on how the subcluster wild bootstrap works, please refer to MacKinnon and Webb (2018).

Results – Bootstrap

Three assumptions must be satisfied for the subcluster wild bootstrap to perform well.

The first assumption – equal cluster sizes and a fixed sample size – is naturally met in this context, as the panel is strongly balanced and the sample size is fixed.

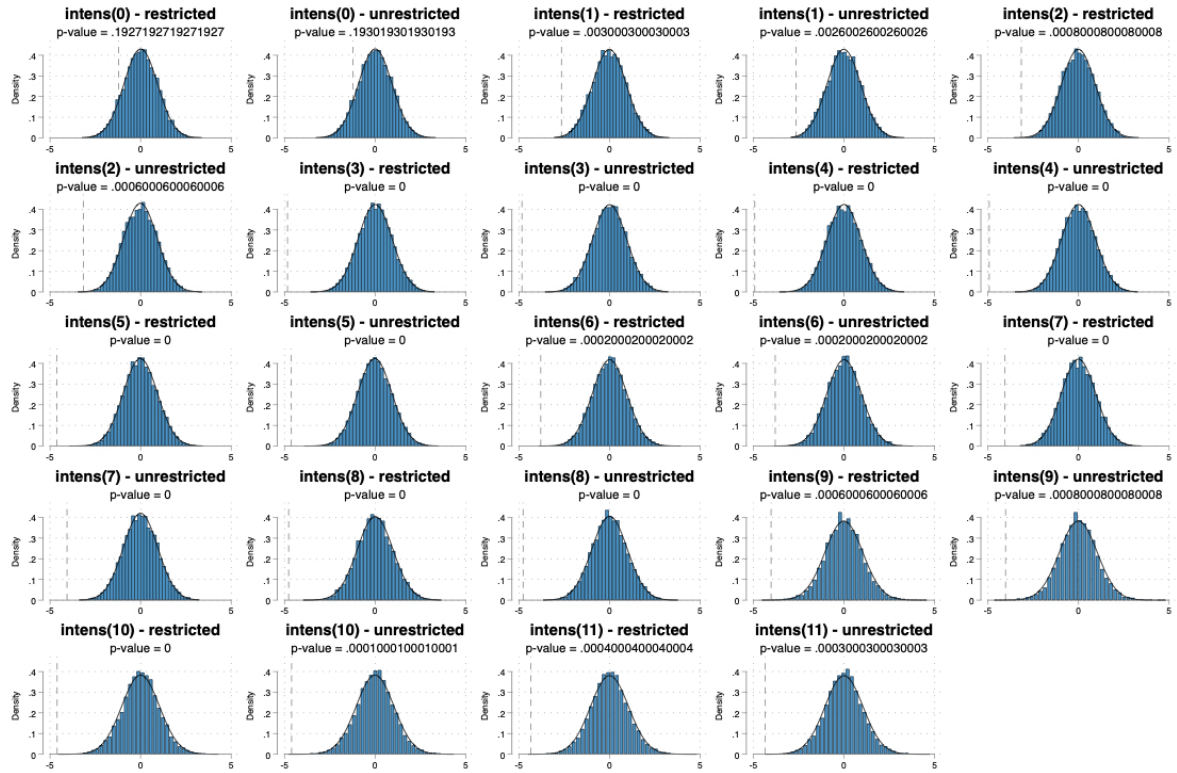
The third assumption – the covariance matrices of each cluster must be proportional – is also likely satisfied here, as this assumption allows for any cross-cluster heteroskedasticity.

The second assumption – that the average within-cluster correlation should be small – is more challenging to check in practice. MacKinnon and Webb (2018) suggest that one way to test this assumption is by comparing the p-values from the restricted and unrestricted bootstrap. Specifically, the p-value from the restricted subcluster wild bootstrap should always be larger than that from the unrestricted bootstrap.

To examine whether this condition holds in this context, Figure D.1 shows the bootstrap distribution of the t-statistic for each value of $w_{g,t} \cdot intens(s)$, both under the null hypothesis and without imposing it. For brevity, only the results for the coefficients of $w_{g,t} \cdot intens(s)$, estimated using (1.5) with the log of current health expenditure as the dependent variable, are shown. Similar results were obtained for the coefficients of (1.2). These results, along with those for the other indicators, are available upon request.

The results indicate that the bootstrap p-values for the restricted and unrestricted subcluster wild bootstrap are very close, suggesting that assumption 2 is not seriously violated in this case.

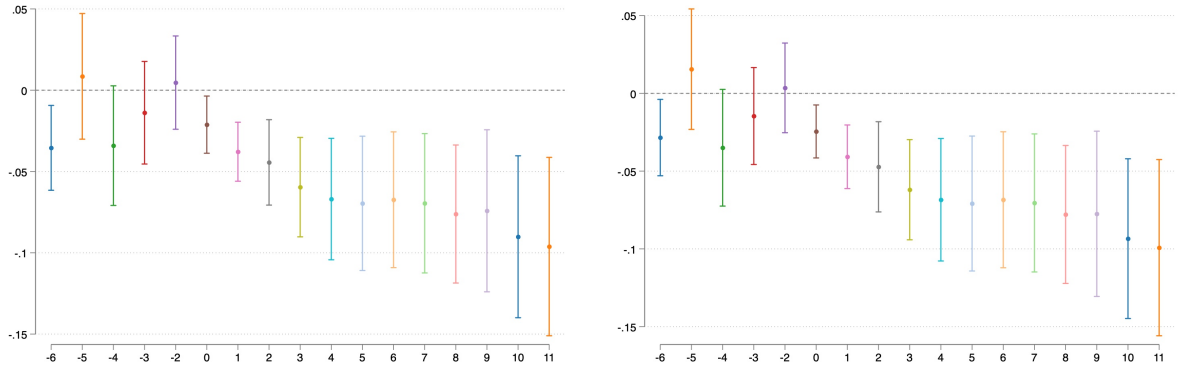
Figure D.1: Bootstrap distribution t-statistics – ln(Current health exp)



Notes: The above histograms report the bootstrap distribution of the t-statistic for each of the coefficients for $w_{g,t} \cdot intens(s)$ obtained by estimating (1.5) using as dependent variable the log of current health expenditure. In particular, these bootstrap distributions were obtained by exploiting the subcluster wild bootstrap with Rademacher weights with $9,999$ replications. Moreover, for each coefficient, the bootstrap distribution obtained by imposing (restricted) and not imposing the null hypothesis (unrestricted) were reported one next to the other.

Appendix E – Additional Results TWM and S&A

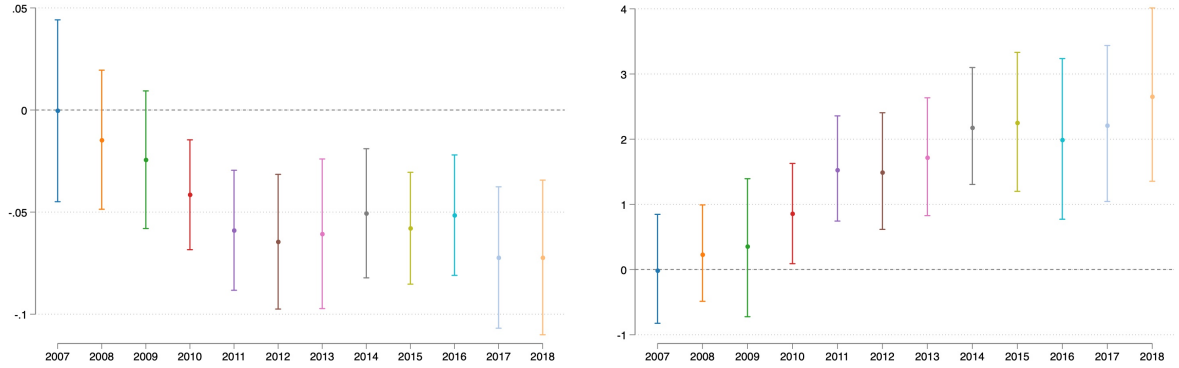
Figure E.1: Results Two-Way Mundlak – $ln(\text{Current health exp})$ with cluster-robust s.e.



(a) Without controls (b) With controls

Notes: The above regressions include 357 RHS-year observations. Results depicted in Panel (a) were obtained by estimating (1.2) using as the dependent variable the log of current health expenditure. In Panel (b) are shown the results obtained by estimating (1.5). Coefficients are reported with 95% confidence intervals. Standard errors are clustered at the RHS level to account for the potential serial correlation of the error term.

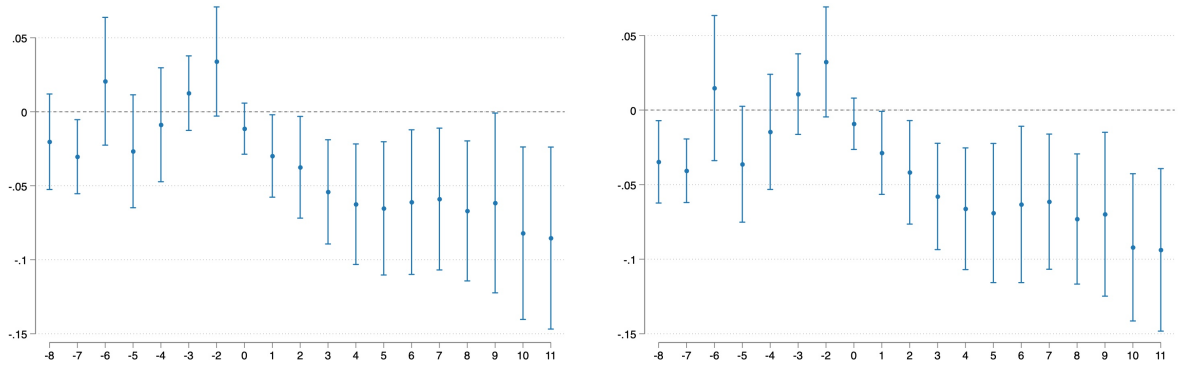
Figure E.2: Results Two-Way Mundlak – ATT that varies by calendar time



(a) $ln(\text{Current health exp})$ (b) Mortality rate from ischaemic heart diseases

Notes: Results depicted in Panel (a) were obtained by estimating (1.11) using as the dependent variable the log of current health expenditure. Panel (b) reports $\hat{\tau}_s$ obtained by estimating (1.11) using as the dependent variable the mortality rate from ischaemic heart diseases. Coefficients, in both Panels, are reported with 95% confidence intervals obtained via the subcluster wild bootstrap (MacKinnon and Webb, 2018) with Rademacher weights. Specifically, the t-statistic is obtained through a CRVE estimator (where the level of clustering is at the RHS level), whereas the resampling is carried out at the RHS-year level.

Figure E.3: Results Sun and Abraham – ln(Current health exp)

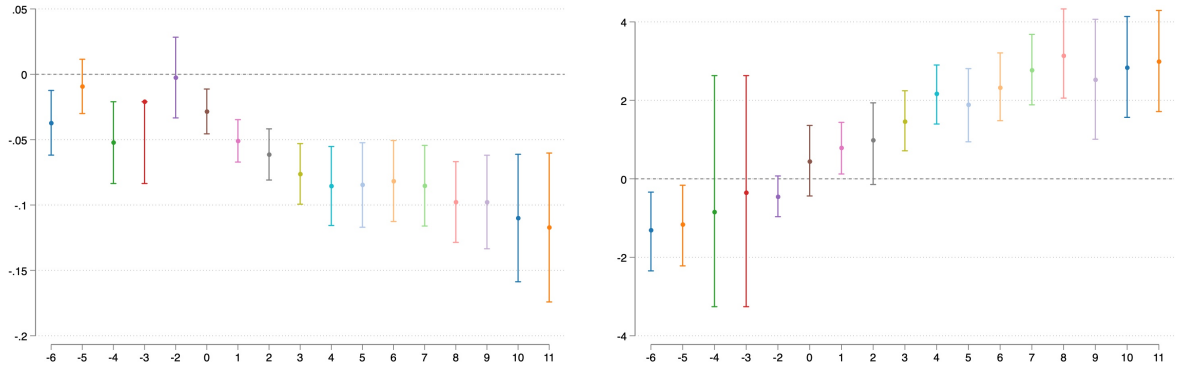


(a) Without controls

(b) With controls

Notes: The above regressions include 357 RHS-year observations. Results depicted in Panel (a) were obtained by exploiting the estimator proposed by Sun and Abraham (2021) – by not conditioning on covariates – using the log of current health expenditure as the dependent variable. Panel (b) are shown the results obtained using the estimator proposed by Sun and Abraham (2021) by conditioning on covariates. Coefficients are reported with 95% confidence intervals. Standard errors are clustered at the RHS level to account for the potential serial correlation of the error term.

Figure E.4: Results Two-Way Mundlak – Excluding Leavers



(a) ln(Current health exp)

(b) Mortality rate from ischaemic heart diseases

Notes: This table reports the results of the TWM obtained by excluding regions that left the treatment status over the observed period. Specifically, results depicted in Panel (a) were obtained by estimating (1.11) using as the dependent variable the log of current health expenditure. Panel (b) reports $\hat{\tau}_s$ obtained by estimating (1.11) using as the dependent variable the mortality rate from ischaemic heart diseases. Coefficients, in both Panels, are reported with 95% confidence intervals obtained via the subcluster wild bootstrap (MacKinnon and Webb, 2018) with Rademacher weights. Specifically, the t-statistic is obtained through a CRVE estimator (where the level of clustering is at the RHS level), whereas the resampling is carried out at the RHS-year level.

Table E.1: Summary Results Two-way Mundlak

	ln(Current health exp)				Hospital beds (ordinary regime)				Hospitalization rate				Inefficiency ratio				Mortality rate from ischaemic heart diseases			
	ATT	95% CI		$p-$value	ATT	95% CI		$p-$value	ATT	95% CI		$p-$value	ATT	95% CI		$p-$value	ATT	95% CI		$p-$value
Short-term	-0.04	[-0.05	-0.02]	0.00	-998.72	[-2051.50	59.82]	0.06	-8.10	[-15.22	-0.82]	0.03	-0.04	[-0.07	-0.00]	0.04	0.58	[-0.05	1.25]	0.07
Long-Term	-0.09	[-0.13	-0.05]	0.00	-1918.43	[-3721.70	-91.75]	0.04	-28.21	[-41.84	-14.61]	0.00	-0.10	[-0.14	-0.05]	0.00	2.46	[1.20	3.72]	0.00
Overall	-0.06	[-0.08	-0.04]	0.00	-1675.03	[-3021.54	-339.75]	0.02	-19.80	[-29.42	-10.35]	0.00	-0.07	[-0.11	-0.03]	0.00	1.68	[0.89	2.49]	0.00
Test CPT	5.88				2.63				1.46				0.99				2.20			
$p-$value Test CPT	0.00				0.03				0.28				0.66				0.35			
Observations	357				441				441				378				336			
Adj. R-squared	0.85				0.93				0.87				0.87				0.82			

Notes: This table reports the estimated short-, long-term, and overall ATT obtained by estimating (1.5) using one of the five indicators above as the dependent variable. Specifically, *Short-term* denotes the average effect over the first three years in which a region was under PdR (i.e., the average of τ_s over $s \in \{0, 1, 2\}$). *Long-term* represents the average effect over the last three observed years, when the RHS is 10 to 12 years after the PdR was first signed. *Overall* represents the average effect of PdRs across all (post-treatment) periods. Coefficients are reported with 95% confidence intervals obtained via the subcluster wild bootstrap with Rademacher weights, and the corresponding $p-$value. Lastly, *Test CPT* refers to the test of pre-trends, conducted by testing the overall significance of τ_s for $s \in \{-6, \ldots, -2\}$ and *$p-$value Test CPT* represents the $p-$value of this test, also obtained via the subcluster wild bootstrap.

Table E.2: Additional Summary Results Two-way Mundlak

	Hospitalization rate – acute				% Patients with hip fracture replacement				% C-sections				Mortality rate				% Patients migrating for ordinary acute hospit.			
	ATT	95% CI		$p-$value	ATT	95% CI		$p-$value	ATT	95% CI		$p-$value	ATT	95% CI		$p-$value	ATT	95% CI		$p-$value
Short-term	-7.78	[-14.70	-0.63]	0.04	0.28	[-6.27	6.95]	0.93	2.53	[0.29	4.94]	0.03	1.10	[-0.55	2.82]	0.18	0.57	[-0.45	1.59]	0.28
Long-Term	-27.25	[-40.11	-14.40]	0.00	2.33	[-12.12	17.07]	0.74	1.99	[-1.90	5.95]	0.30	4.53	[0.09	9.14]	0.05	3.47	[1.85	5.11]	0.00
Overall	-18.93	[-27.90	-10.05]	0.00	-0.14	[-8.11	7.81]	0.97	2.37	[-0.21	5.14]	0.07	2.94	[0.87	4.99]	0.01	2.06	[0.94	3.17]	0.00
Test CPT	1.54				1.36				2.28				16.19				1.31			
$p-$value Test CPT	0.27				0.27				0.16				0.02				0.24			
Observations	441				378				441				336				420			
Adj. R-squared	0.89				0.81				0.94				0.94				0.79			

Notes: This table reports the estimated short-, long-term, and overall ATT obtained by estimating (1.5) using one of the five indicators above as the dependent variable. Specifically, *Short-term* denotes the average effect over the first three years in which a region was under PdR (i.e., the average of τ_s over $s \in \{0, 1, 2\}$). *Long-term* represents the average effect over the last three observed years, when the RHS is 10 to 12 years after the PdR was first signed. *Overall* represents the average effect of PdRs across all (post-treatment) periods. Coefficients are reported with 95% confidence intervals obtained via the subcluster wild bootstrap with Rademacher weights, along with the corresponding $p-$value. Lastly, *Test CPT* refers to the test of pre-trends, conducted by testing the overall significance of τ_s for $s \in \{-6, \ldots, -2\}$ and *$p-$value Test CPT* represents the $p-$value of this test, also obtained via the subcluster wild bootstrap.

Appendix F – Additional Results Bounds

Table F.1: Bounds on treatment effect based on ($\delta_{T,Max}$, $\delta_{SC,Max}$)

	Abruzzo		Calabria		Campania		Lazio		Liguria		Molise		Piemonte		Puglia		Sardegna		Sicilia	
	Hospitalization rate (acute)																			
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2007	-23.03	16.20			-4.52	14.26	-11.81	-5.15	-24.63	17.17	-24.56	6.17					-22.69	11.17	-27.78	9.22
2008	-39.48	-1.60			-4.46	13.44	-11.56	-6.81	-25.56	16.24	-34.21	-4.43					-24.78	8.11	-33.81	3.19
2009	-56.60	-15.48			-6.29	8.70	-14.65	-12.31	-36.38	5.42	-48.17	-15.12					-32.36	-2.53	-45.50	-8.70
2010	-62.62	-22.41	-21.56	-1.29	-15.06	-2.49			-39.63	2.17	-41.78	-9.20	-2.64	10.56			-29.67	-2.30	-43.43	-8.86
2011	-64.46	-27.97	-29.42	-4.70	-16.61	-7.26			-51.81	-10.01	-55.49	-25.53	-4.04	8.80	-22.73	1.73	-27.34	-3.09	-47.52	-16.88
2012	-64.14	-28.94	-35.84	-10.00	-17.73	-10.19			-39.65	2.15	-65.99	-37.17	-5.86	7.34	-29.91	-4.02	-29.55	-7.17	-47.84	-18.58
2013	-63.41	-30.72	-44.77	-18.06	-20.64	-14.55			-40.17	1.63	-64.70	-37.70	-7.12	3.27	-33.84	-6.81	-34.21	-13.32	-52.31	-24.20
2014	-68.00	-35.56	-52.66	-26.61	-23.20	-17.05			-42.04	-0.24	-81.52	-54.22	-8.93	0.36	-45.63	-18.99	-35.29	-14.39	-57.92	-29.57
2015	-68.63	-36.41	-56.75	-31.75	-21.64	-15.02			-41.94	-0.14	-68.32	-41.80	-9.65	0.21	-38.79	-12.94	-36.64	-15.32	-60.67	-31.51
2016	-66.38	-36.27	-54.96	-27.90	-22.08	-18.54			-42.06	-0.26	-68.60	-46.30	-12.58	-1.60	-41.42	-13.22	-37.56	-19.39	-61.44	-34.65
2017	-66.73	-38.19	-53.42	-23.98	-23.87	-22.63			-34.38	7.42	-68.84	-49.15	-14.41	-3.40	-42.73	-14.53	-37.82	-21.94	-59.67	-35.20
2018	-66.26	-39.11	-53.02	-23.02					-34.37	7.43	-72.40	-55.21	-14.56	-3.04	-42.05	-13.85	-9.92	4.12	-59.50	-36.81
	Mortality rate (all causes)																			
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2007	-0.54	2.26			-2.55	4.17	-1.69	-0.29	-6.63	0.17	-4.39	-0.59					-3.92	2.28	-3.01	5.19
2008	-2.28	0.52			-1.79	5.01	-1.78	-0.38	-6.54	0.26	-1.11	2.27					-2.02	4.18	-2.45	5.75
2009	1.77	4.57			-1.19	5.09	-0.85	0.55	-4.44	2.36	-3.22	1.19					0.61	6.81	-1.18	7.02
2010	1.59	4.39	-2.42	2.78	-1.32	5.48	-0.15	1.25	-5.12	1.68	-1.87	2.40	-2.94	1.06			-2.35	3.85	-3.39	4.81
2011	1.25	3.01	-3.58	1.62	1.29	3.34	1.54	2.94	-3.24	3.56	3.92	5.21	-2.14	1.86	-2.21	4.47	0.18	4.96	-0.29	4.30
2012			-0.84	2.66			2.40	3.80	-5.28	1.52			-1.58	2.42	0.65	3.91	4.06	4.86	3.34	4.49
2013	2.52	2.87	-3.78	0.95			0.31	1.71	-2.16	4.64			-0.48	3.52	-2.95	1.54	1.21	4.04	-2.01	0.39
2014	0.56	2.02	-1.99	2.64			-1.22	0.18	-5.96	0.84	2.47	2.77	0.52	4.52	-1.64	2.75	1.83	4.23	-1.70	0.59
2015									-4.80	2.00			1.75	5.75						
2016							-1.56	-0.16	-6.21	0.59			0.82	4.82						
2017									-2.74	4.06			2.56	6.56						
2018									0.61	7.41			5.05	9.05						
	% patients (aged 65+) w/ hip fracture operated on $\leq$ 48 hrs																			
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2007	-2.77	2.51			-3.30	0.70	1.18	4.10	-9.11	4.06	-11.57	3.83					1.48	10.04	-2.18	1.04
2008	-1.69	2.00			-3.23	-0.42	3.93	5.92	-11.75	-4.07	-7.24	8.16					1.06	8.55	-1.09	0.38
2009					-5.59	-1.59	5.44	7.95	-4.37	-2.78	-10.05	4.95					-4.69	5.89	-3.12	0.28
2010			-6.22	-1.95	-4.88	-0.88	10.20	14.20			-9.61	3.88	0.45	2.98			-6.10	5.21	-3.94	-0.54
2011			-5.53	-1.52	-4.44	-0.44	14.41	18.41			-12.52	2.11			3.41	7.68	-2.76	9.15	4.08	7.48
2012							20.05	22.79			-21.05	-14.62					-5.83	5.88		
2013																	-0.13	0.66		
2014																	8.40	9.37		
2015																				
2016																				
2017																				
2018																				

Table F.2: Bounds on treatment effect based on ($\delta_{T,p_{75}}$, $\delta_{SC,p_{75}}$)

	Abruzzo		Calabria		Campania		Lazio		Liguria		Molise		Piemonte		Puglia		Sardegna		Sicilia	
	Mortality rate from ischaemic heart diseases																			
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2007	-0.74	0.26			-0.57	0.43	-0.62	-0.02	-1.23	-0.23	-0.52	1.28					-0.10	0.10	-0.15	0.65
2008	-0.31	0.69			-0.31	0.69	-0.35	0.25	-0.39	0.61	1.02	2.82					0.19	0.39	0.08	0.88
2009	0.88	1.88			0.51	1.42	-0.72	-0.12	-0.18	0.82	-1.05	0.74							0.82	1.25
2010	0.50	1.50	-1.12	0.01	0.21	1.21	-0.20	0.40	0.20	1.20	0.32	2.12							0.06	0.86
2011	1.02	2.02	-0.32	0.32	0.20	1.03	0.37	0.97	1.05	2.05	1.72	3.52	0.37	0.40	0.23	0.88	0.54	0.74	0.09	0.89
2012	1.04	1.99					-0.36	0.24	1.91	2.91	0.10	1.90					-0.52	-0.32		
2013	1.49	2.49	-0.64	-0.54	-0.10	0.19	-0.16	0.44	1.72	2.72	1.11	2.91			0.03	0.21			-0.75	0.03
2014	1.93	2.64	0.08	0.93	-0.31	0.69	-0.36	-0.00	1.63	1.91	2.66	4.40			-0.04	0.93			-0.42	0.38
2015	2.50	3.50	0.07	0.37	0.86	1.35	-0.26	0.34	0.79	1.70	3.38	5.18			0.77	1.19			-0.53	0.27
2016	1.96	2.75	-0.73	-0.43	-0.40	0.09					0.06	1.70			-0.50	0.06			-1.35	-0.55
2017	1.91	2.77	0.22	0.24	-0.33	-0.12	-0.38	-0.25			1.89	3.69			0.33	0.58			-1.17	-0.37
2018			-0.62	0.45	-0.34	0.66									0.00	0.92			-0.26	-0.18

Table F.3: Reasonable values of δ

	Abruzzo		**Calabria**		**Campania**		**Lazio**		**Liguria**		**Molise**		**Piemonte**		**Puglia**		**Sardegna**		**Sicilia**	
	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}
								ln(**current health expenditure**)												
2002		0.0		0.0		-0.0		-0.1		-0.0		-0.1		-0.0		0.0		0.0		-0.0
2003	0.1	0.0	0.0	0.0	0.0	-0.0	0.1	-0.0	0.0	-0.0	0.2	0.0	0.0	-0.0	0.0	-0.0	0.0	-0.0	0.0	-0.0
2004	-0.0	-0.1	0.1	0.0	0.1	0.0	0.2	0.1	0.1	0.0	-0.0	-0.1	0.1	0.0	0.1	-0.0	0.1	-0.0	0.1	0.0
2005	0.1	0.0	0.0	-0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.2	0.1	0.0	-0.0	0.1	0.0	0.1	0.0	0.0	-0.0
2006	-0.0	-0.0	0.1	-0.0	-0.0	-0.0	0.1	0.0	0.0	-0.0	-0.1	-0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	0.1	0.0
2007			0.1	0.0									0.0	-0.0	0.1	0.0				
2008			0.0	-0.0									0.0	-0.0	0.0	0.0				
2009			0.0	0.0									0.0	-0.0	0.0	0.0				
2010															0.0	0.0				
Max	0.1	0.1	0.1	0.0	0.1	0.0	0.2	0.1	0.1	0.0	0.2	0.1	0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.0
p75	0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.1	0.1	0.0	0.2	0.1	0.0	0.0	0.1	0.0	0.1	0.0	0.1	0.0
								Hospitalization rate												
	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}
1998		7.7		-2.7		-13.4		-4.2		7.1		-6.7		-2.0		-0.8		-13.8		-9.3
1999	-26.1	-13.6	-16.8	-11.1	-3.4	-11.5	-8.7	-9.0	-1.3	7.2	-18.9	-29.4	-2.8	0.8	-13.1	-6.8	-19.5	-28.4	-18.6	-20.5
2000	-6.7	-14.6	-7.7	-11.8	-6.9	-10.0	-4.6	-7.1	-1.8	9.1	4.6	-18.0	-11.2	-2.9	-4.3	-3.5	3.0	-16.8	-13.6	-26.9
2001	-4.7	-14.3	1.6	-5.2	1.2	-4.5	0.8	-5.8	-1.4	12.2	7.2	-1.2	-2.3	0.5	-4.0	-2.7	-5.6	-18.1	15.7	-6.6
2002	8.1	2.1	-6.0	-0.4	-5.8	-1.1	-3.6	-1.4	-15.8	1.7	-9.4	-1.9	-11.6	-6.2	-8.1	-0.8	8.2	-0.9	7.4	11.0
2003	-20.6	6.7	-18.6	14.5	-18.5	7.9	-8.4	4.8	-37.6	-20.9	-12.6	9.2	-4.7	-0.3	-17.0	13.4	-2.8	23.4	-16.0	26.1
2004	-26.3	-19.4	-3.5	10.2	-2.0	7.3	-0.1	9.0	1.4	-18.0	4.8	12.9	-1.8	5.7	-9.9	3.5	0.5	25.6	-9.1	16.9

(Continues)

Table F.3 (Continued)

	Abruzzo		Calabria		Campania		Lazio		Liguria		Molise		Piemonte		Puglia		Sardegna		Sicilia	
	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}
2005	41.6	24.8	-2.0	8.9	6.8	16.5	-3.5	7.7	14.7	1.6	3.3	20.1	-2.1	4.7	-1.2	3.6	-5.6	22.4	-6.9	11.1
2006	4.0	20.7	4.4	-2.4	3.7	8.8	2.0	6.1	-3.1	-0.1	2.2	15.1	-3.4	-0.3	4.4	-5.9	-4.7	6.7	1.2	-1.7
2007			-11.2	-4.9									-1.3	-0.7	-4.1	-2.3				
2008			-6.1	-11.7									-2.6	-0.9	-5.5	-7.9				
2009			-5.7	-16.9									5.2	5.4	-2.5	-9.0				
2010															-3.0	-10.3				
Max	41.6	24.8	18.6	16.9	18.5	16.5	8.7	9.0	37.6	20.9	18.9	29.4	11.6	6.2	17.0	13.4	19.5	28.4	18.6	26.9
p75	26.2	19.4	11.2	11.8	6.9	11.5	6.5	7.7	15.2	12.2	11.0	18.0	5.2	5.1	9.0	7.9	6.9	23.4	15.8	20.5
	Hospitalization rate (Acute)																			
	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}
1998		9.1		-2.1		-14.2		-4.6		7.9		-4.8		-1.5		0.6		-14.4		-8.7
1999	-26.8	-13.4	-16.9	-11.5	-2.4	-11.9	-8.0	-9.4	-1.0	8.1	-19.7	-28.6	-4.2	0.8	-12.8	-6.0	-19.1	-29.0	-18.5	-20.9
2000	-4.3	-11.8	-7.9	-12.2	-7.5	-10.8	-4.6	-7.2	-2.0	9.7	4.2	-17.2	-10.4	-1.8	-5.4	-3.6	3.0	-17.4	-13.7	-27.0
2001	-4.6	-11.1	1.5	-5.4	1.1	-5.0	0.3	-6.0	-1.7	12.7	6.7	-0.6	-3.7	0.8	-4.0	-2.5	-5.7	-18.4	15.7	-6.3
2002	5.0	2.5	-6.2	-0.8	-5.8	-1.5	-3.9	-1.7	-16.8	1.6	-9.7	-1.3	-12.2	-6.3	-8.1	-0.5	8.2	-1.0	6.7	10.6
2003	-20.5	7.0	-18.4	14.2	-18.1	8.3	-8.4	4.7	-37.5	-20.9	-14.6	7.8	-4.9	-1.0	-16.4	14.1	-3.0	23.6	-16.0	25.5
2004	-27.2	-19.6	-3.2	10.3	-2.3	7.6	-0.3	8.9	1.0	-18.2	6.6	12.8	-2.2	4.8	-11.6	2.8	0.5	25.8	-9.2	16.5
2005	38.5	22.1	-2.3	9.6	6.9	17.5	-3.6	8.0	13.3	0.5	1.8	19.2	-2.0	4.1	-2.5	2.5	-5.7	23.3	-7.1	11.4
2006	1.2	15.3	4.0	-2.0	3.7	9.9	2.1	7.2	-3.3	-1.3	0.5	12.7	-2.2	0.2	3.7	-7.5	-4.7	7.5	1.0	-1.2
2007			-11.0	-3.9									-1.5	0.2	-4.2	-3.6				
2008			-5.8	-10.1									-2.2	0.5	-5.5	-8.9				
2009			-5.8	-15.0									4.6	6.6	-2.4	-9.5				
2010															-2.6	-9.9				

(Continues)

Table F.3 (Continued)

	Abruzzo		Calabria		Campania		Lazio		Liguria		Molise		Piemonte		Puglia		Sardegna		Sicilia	
	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}
Max	38.5	22.1	18.4	15.0	18.1	17.5	8.4	9.4	37.5	20.9	19.7	28.6	12.2	6.6	16.4	14.1	19.1	29.0	18.5	27.0
p75	27.0	15.3	11.0	11.9	7.2	11.9	6.3	8.0	15.0	12.7	12.1	17.2	4.9	4.4	9.9	8.9	6.9	23.6	15.9	20.9
										Hospital beds (ordinary)										
	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}
1998		1438.7		741.6		5447.4		1529.5		-654.2		-343.3		-654.9		4285.1		873.3		1782.1
1999	-1759.0	199.6	-1205.0	-232.4	-5982.0	385.9	-3637.0	555.6	-1279.0	-435.2	-152.0	-311.7	-396.0	797.8	-2572.0	1935.2	-858.0	198.3	-2380.0	125.8
2000	-1162.0	-812.9	66.0	-171.4	-995.0	-416.5	-2047.0	-762.7	-53.0	-202.2	-130.0	-338.0	-835.0	589.0	-1393.0	543.4	202.0	428.4	-622.0	-376.6
2001	-194.0	-849.8	1097.0	939.6	18.0	-147.8	-534.0	-467.4	-184.0	141.8	4.0	-269.8	-598.0	448.5	-687.0	-127.4	-762.0	-307.4	-114.0	-317.7
2002	584.0	-271.5	-1775.0	-796.4	-452.0	-188.3	-951.0	-55.4	-148.0	-315.2	50.0	-160.5	-1848.0	-1005.5	-845.0	-934.1	7.0	-265.4	427.0	406.7
2003	-593.0	-59.2	1716.0	1426.6	-2583.0	-2125.8	-1233.0	-211.2	-1994.0	-304.2	86.0	310.8	-195.0	19.0	-564.0	-1015.8	-275.0	-166.9	-353.0	703.5
2004	-657.0	-654.5	-1547.0	-126.4	472.0	-1541.2	-356.0	-95.2	117.0	-178.2	-8.0	286.7	-269.0	149.2	-986.0	-2002.1	-104.0	-245.9	-1099.0	-330.6
2005	929.0	381.5	-725.0	-813.4	831.0	-698.1	-123.0	-267.8	854.0	881.8	72.0	449.8	-364.0	-176.8	565.0	-1396.0	-14.0	-205.2	-225.0	-541.4
2006	215.0	628.2	-7.0	-967.4	103.0	-715.6	92.0	-225.3	-239.0	1065.8	-22.0	375.9	-217.0	-166.4	247.0	-1288.4	2.0	-309.2	-767.0	-1451.7
2007			-208.0	-1090.4									-36.0	28.7	-154.0	-1358.2				
2008			-810.0	-1806.4									-126.0	125.8	-370.0	-1640.0				
2009			-291.0	-2088.4									182.0	467.0	-192.0	-1822.6				
2010															-15.0	-1858.3				
Max	1759.0	1438.7	1775.0	2088.4	5982.0	5447.4	3637.0	1529.5	1994.0	1065.8	152.0	449.8	1848.0	1005.5	2572.0	4285.1	858.0	873.3	2380.0	1782.1
p75	1045.5	812.9	1547.0	1258.5	1789.0	1541.2	1640.0	555.6	1066.5	654.2	108.0	343.3	598.0	622.0	915.5	1858.3	518.5	309.2	933.0	703.5
										Inefficiency Ratio										
2001		-0.0		0.0		-0.0		0.0		0.0		0.0		0.0		-0.0		-0.0		0.0

(Continues)

Table F.3 (Continued)

	Abruzzo		Calabria		Campania		Lazio		Liguria		Molise		Piemonte		Puglia		Sardegna		Sicilia	
	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}
2002	-0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.1	0.0	-0.1	-0.0	-0.0	0.0	-0.1	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0
2003	-0.1	-0.0	-0.1	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.1	-0.0
2004	0.0	0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.0
2005	-0.0	0.0	-0.0	0.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.1	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	0.0	-0.0	0.0
2006	-0.0	0.0	-0.0	0.0	0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	0.0	-0.0	0.0
2007			-0.0	0.0									-0.0	-0.0	-0.0	0.0				
2008			-0.0	-0.0									-0.0	-0.0	-0.0	0.0				
2009			0.2	0.0									0.2	0.0	0.3	0.1				
2010															-0.0	0.1				
Max	0.1	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.1	0.0	0.2	0.0	0.3	0.1	0.0	0.0	0.1	0.0
p75	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	% C-sections																			
	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}
1998		-3.0		0.7		-4.1		0.6		-2.1		-2.3		0.8		-0.3		2.1		0.2
1999	2.8	-2.6	1.3	-2.5	3.7	-4.3	-0.6	-0.5	1.2	-0.9	1.6	-2.7	0.3	1.3	2.2	-2.5	0.8	-0.9	1.6	-2.4
2000	5.1	2.3	5.3	4.0	5.0	1.5	2.1	0.2	2.1	-0.7	0.3	-2.9	-1.6	-2.2	3.4	2.0	1.1	1.0	0.5	-0.9
2001	0.7	2.6	0.9	4.5	2.4	5.0	-2.6	-1.2	0.5	-0.9	2.8	-0.4	0.8	-0.9	3.0	4.8	-2.5	-0.5	4.3	3.7
2002	-0.9	-1.0	-0.7	-1.9	0.9	0.2	3.6	0.0	0.6	1.3	3.5	0.0	1.5	0.3	-0.1	-1.0	5.4	-0.8	-0.5	-2.5
2003	3.1	-1.0	3.1	-3.2	2.1	-0.8	1.1	-0.1	0.8	0.4	1.1	-1.9	0.2	-0.6	2.5	-2.9	0.8	-3.1	3.3	-3.1
2004	1.1	-0.7	1.0	-2.6	1.8	-0.3	-0.0	-0.9	1.2	0.6	1.9	-1.3	1.3	0.0	0.5	-2.9	3.4	-1.0	2.8	-1.0
2005	0.8	0.4	2.2	0.5	0.9	1.1	1.8	0.5	-0.0	-0.1	6.9	5.6	1.7	1.1	2.5	0.5	2.6	2.1	2.4	2.1
2006	2.5	3.0	-0.1	0.5	0.9	1.7	1.7	1.4	2.4	2.4	-0.3	5.8	-0.5	0.3	1.8	2.3	-0.5	1.3	1.9	3.9
2007			1.8	4.9									0.5	0.9	0.7	5.6				

(Continues)

Table F.3 (Continued)

	Abruzzo		Calabria		Campania		Lazio		Liguria		Molise		Piemonte		Puglia		Sardegna		Sicilia	
	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}
2008			-0.5	5.2									0.2	0.9	0.7	7.0				
2009			1.0	6.7									-0.2	1.3	-1.3	6.3				
2010															-0.9	5.4				
Max	5.1	3.0	5.3	6.7	5.0	5.0	3.6	1.4	2.4	2.4	6.9	5.8	1.7	2.2	3.4	7.0	5.4	3.1	4.3	3.9
p75	3.0	2.6	2.2	4.7	3.0	4.1	2.4	0.9	1.7	1.3	3.2	2.9	1.5	1.2	2.5	5.4	3.0	2.1	3.1	3.1
										‰ Mortality rate										
	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}
2003		1.4		2.3		3.4		-0.1		3.4		-1.2		0.5		3.4		3.1		4.1
2004	-6.0	0.8	-5.3	-0.9	-6.2	-0.6	-7.6	0.7	-16.1	-1.6	-0.3	2.8	-12.2	-2.0	-6.4	-0.8	-7.0	-0.9	-7.4	-1.1
2005	2.6	-0.8	6.0	0.9	4.4	-0.4	1.5	-0.2	4.1	-1.8	2.1	-1.3	4.9	0.9	3.5	-1.5	3.2	-0.6	4.0	-1.3
2006	-2.7	-1.4	-3.9	-2.3	-2.8	-2.5	-2.3	-0.4	-2.0	-0.1	-2.5	-0.3	-2.7	0.6	-0.4	-1.1	-2.2	-1.7	-1.3	-1.8
2007			3.4	0.7									-0.3	-0.5	3.8	2.2				
2008			0.1	2.1									2.3	-0.6	-1.5	2.0				
2009			2.2	2.6									0.6	1.1	0.9	1.2				
2010															-0.0	1.7				
Max	6.0	1.4	6.0	2.6	6.2	3.4	7.6	0.7	16.1	3.4	2.5	2.8	12.2	2.0	6.4	3.4	7.0	3.1	7.4	4.1
p75	6.0	1.4	5.3	2.3	6.2	2.9	7.6	0.6	16.1	2.6	2.5	2.0	4.9	1.1	3.8	2.1	7.0	2.4	7.4	3.0
										‰ Mortality rate ischaemic heart diseases										
2003		0.6		0.8		0.5		0.2		0.7		-0.0		-0.5		0.8		-0.0		0.5
2004	-1.2	-0.1	-1.1	-0.4	-0.9	-0.5	-1.4	0.2	-2.0	-0.0	-1.5	-1.1	-1.0	0.0	-1.4	-0.5	-0.6	-0.1	-0.9	-0.2
2005	0.2	-0.4	0.6	-0.2	1.0	0.1	-0.0	-0.0	0.0	-0.3	2.4	0.5	0.4	0.1	0.7	-0.3	0.2	0.1	0.3	-0.4

(Continues)

Table F.3 (Continued)

	Abruzzo		Calabria		Campania		Lazio		Liguria		Molise		Piemonte		Puglia		Sardegna		Sicilia	
	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}
2006	-0.6	-0.0	-0.7	-0.3	-0.8	-0.1	-1.0	-0.4	-1.2	-0.4	-1.4	0.7	-0.5	0.4	-0.5	-0.0	-0.5	0.0	-0.3	0.2
2007			-0.1	-0.6									-0.2	0.8	0.6	0.3				
2008			0.6	0.2									0.4	1.4	-0.7	-0.1				
2009			-0.3	0.6									-0.6	1.3	0.4	0.9				
2010															-0.5	-0.1				
Max	1.2	0.6	1.1	0.8	1.0	0.5	1.4	0.4	2.0	0.7	2.4	1.1	1.0	1.4	1.4	0.9	0.6	0.1	0.9	0.5
p75	1.2	0.5	0.7	0.6	1.0	0.5	1.4	0.3	2.0	0.5	2.4	0.9	0.6	1.3	0.7	0.7	0.6	0.1	0.9	0.4
	% Patients migrating for ordinary acute hospit.																			
	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}
1999		0.4		-0.9		-0.1		-0.5		-1.1		0.9		-0.4		-0.9		-0.2		0.5
2000	0.1	0.2	-0.1	-1.1	-0.1	-0.2	0.5	0.1	0.4	-0.6	0.1	0.4	0.3	-0.1	-0.2	-1.2	0.2	-0.1	0.4	0.8
2001	0.1	0.4	0.2	-0.4	0.5	0.7	0.2	0.3	0.5	-0.2	-1.0	-0.6	0.2	-0.2	0.3	-0.4	0.8	0.9	-0.9	0.4
2002	-0.7	-0.6	0.3	-0.5	-0.2	0.2	0.2	0.3	0.8	0.2	0.2	-0.5	0.6	0.2	0.3	-0.5	-0.8	-0.2	-0.6	-0.6
2003	-0.0	-0.9	0.6	-0.1	-0.1	-0.2	-0.1	0.0	0.6	0.6	1.7	1.0	-0.0	0.3	0.6	-0.1	-0.1	-0.4	0.1	-0.7
2004	0.6	-0.6	0.5	0.0	-0.2	-0.7	0.1	0.0	0.1	0.6	-0.9	-0.3	-0.2	0.2	0.6	0.2	0.2	-0.4	0.1	-0.9
2005	0.5	0.2	0.4	1.3	-0.1	-0.1	-0.1	-0.1	-0.1	0.4	-0.3	-0.4	-0.0	0.0	0.2	1.2	0.2	-0.2	0.1	-0.1
2006	0.5	0.9	-0.2	1.7	0.0	0.5	0.1	-0.0	-0.0	0.2	-0.3	-0.4	-0.1	-0.2	-0.1	1.8	0.6	0.6	-0.0	0.5
2007			1.0	1.7									-0.3	-0.5	-0.3	0.5				
2008			0.5	2.8									-0.1	-0.6	-0.1	0.9				
2009			1.2	4.7									-0.9	-1.3	0.0	1.7				
2010															-0.2	1.4				
Max	0.7	0.9	1.2	4.7	0.5	0.7	0.5	0.5	0.8	1.1	1.7	1.0	0.9	1.3	0.6	1.8	0.8	0.9	0.9	0.9
p75	0.6	0.7	0.6	1.7	0.2	0.6	0.2	0.3	0.6	0.6	1.0	0.7	0.3	0.5	0.3	1.3	0.8	0.5	0.6	0.7

(Continues)

Table F.3 (Continued)

	Abruzzo		Calabria		Campania		Lazio		Liguria		Molise		Piemonte		Puglia		Sardegna		Sicilia	
	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}
	% patients (aged 65+) w/ hip fracture operated on ≤ 48 hrs																			
	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}	δ_T	δ_{SC}
2001		3.2		-0.6		-3.8		-0.4		1.5		4.1		3.1		-2.5		-0.4		-4.3
2002	-2.0	1.4	-1.7	-0.3	2.0	0.1	0.1	-1.2	-0.9	-3.0	-7.0	-0.2	1.3	2.3	2.0	1.5	-3.2	-2.2	-0.9	-3.2
2003	-0.2	2.2	-1.4	2.5	-1.2	2.7	2.0	0.2	-0.3	-7.1	2.3	3.7	3.4	2.6	-2.6	3.1	0.4	-1.8	1.7	2.0
2004	-3.2	-2.5	0.3	1.1	-1.4	0.4	-0.7	-0.4	7.8	0.2	4.0	4.7	-1.2	-0.7	-2.1	-0.6	-5.4	-3.7	0.2	1.6
2005	0.3	-3.2	-0.7	-0.1	0.0	-0.4	-1.1	-1.3	1.5	1.9	-9.1	-7.7	-0.5	-4.8	0.8	-0.3	7.3	1.7	0.7	1.3
2006	1.4	-1.0	-3.0	-2.4	0.3	1.0	1.5	3.2	4.1	6.6	2.0	-4.6	0.4	-2.4	-1.6	-1.2	1.3	6.4	0.1	2.7
2007			-1.2	-3.2									-0.3	-4.0	-2.5	-3.3				
2008			3.0	1.7									-1.3	-3.6	-0.9	-2.3				
2009			0.5	-2.5									0.2	-4.0	-1.0	-8.1				
2010															1.3	-6.2				
Max	3.2	3.2	3.0	3.2	2.0	3.8	2.0	3.2	7.8	7.1	9.1	7.7	3.4	4.8	2.6	8.1	7.3	6.4	1.7	4.3
p75	2.0	3.2	2.4	2.5	1.4	2.7	1.5	1.3	4.1	6.6	7.0	4.7	1.3	4.0	2.1	3.3	5.4	3.7	0.9	3.2

Table F.4: TE for Mortality rate from ischaemic heart diseases – complete set of results for **Abruzzo**

Year	δ_{SC}	**0**		**0.5**		**1**		**1.5**		**2**		**2.5**		**3**		**3.5**		**4**		**4.5**		**5**	
	δ_T	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2007	0.0	0.02	-0.24	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
2007	0.5	-0.24	-0.24	-0.48	0.26	-0.48	0.52	-0.48	0.52	-0.48	0.52	-0.48	0.52	-0.48	0.52	-0.48	0.52	-0.48	0.52	-0.48	0.52	-0.48	0.52
2007	1.0	-0.24	-0.24	-0.74	0.26	-0.98	0.76	-0.98	1.02	-0.98	1.02	-0.98	1.02	-0.98	1.02	-0.98	1.02	-0.98	1.02	-0.98	1.02	-0.98	1.02
2007	1.5	-0.24	-0.24	-0.74	0.26	-1.24	0.76	-1.48	1.26	-1.48	1.52	-1.48	1.52	-1.48	1.52	-1.48	1.52	-1.48	1.52	-1.48	1.52	-1.48	1.52
2007	2.0	-0.24	-0.24	-0.74	0.26	-1.24	0.76	-1.74	1.26	-1.98	1.76	-1.98	2.02	-1.98	2.02	-1.98	2.02	-1.98	2.02	-1.98	2.02	-1.98	2.02
2007	2.5	-0.24	-0.24	-0.74	0.26	-1.24	0.76	-1.74	1.26	-2.24	1.76	-2.48	2.26	-2.48	2.52	-2.48	2.52	-2.48	2.52	-2.48	2.52	-2.48	2.52
2007	3.0	-0.24	-0.24	-0.74	0.26	-1.24	0.76	-1.74	1.26	-2.24	1.76	-2.74	2.26	-2.98	2.76	-2.98	3.02	-2.98	3.02	-2.98	3.02	-2.98	3.02
2007	3.5	-0.24	-0.24	-0.74	0.26	-1.24	0.76	-1.74	1.26	-2.24	1.76	-2.74	2.26	-3.24	2.76	-3.48	3.26	-3.48	3.52	-3.48	3.52	-3.48	3.52
2007	4.0	-0.24	-0.24	-0.74	0.26	-1.24	0.76	-1.74	1.26	-2.24	1.76	-2.74	2.26	-3.24	2.76	-3.74	3.26	-3.98	3.76	-3.98	4.02	-3.98	4.02
2007	4.5	-0.24	-0.24	-0.74	0.26	-1.24	0.76	-1.74	1.26	-2.24	1.76	-2.74	2.26	-3.24	2.76	-3.74	3.26	-4.24	3.76	-4.48	4.26	-4.48	4.52
2007	5.0	-0.24	-0.24	-0.74	0.26	-1.24	0.76	-1.74	1.26	-2.24	1.76	-2.74	2.26	-3.24	2.76	-3.74	3.26	-4.24	3.76	-4.74	4.26	-4.98	4.76
2008	0.0	0.25	0.19	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
2008	0.5	0.19	0.19	-0.25	0.69	-0.25	0.75	-0.25	0.75	-0.25	0.75	-0.25	0.75	-0.25	0.75	-0.25	0.75	-0.25	0.75	-0.25	0.75	-0.25	0.75
2008	1.0	0.19	0.19	-0.31	0.69	-0.75	1.19	-0.75	1.25	-0.75	1.25	-0.75	1.25	-0.75	1.25	-0.75	1.25	-0.75	1.25	-0.75	1.25	-0.75	1.25
2008	1.5	0.19	0.19	-0.31	0.69	-0.81	1.19	-1.25	1.69	-1.25	1.75	-1.25	1.75	-1.25	1.75	-1.25	1.75	-1.25	1.75	-1.25	1.75	-1.25	1.75
2008	2.0	0.19	0.19	-0.31	0.69	-0.81	1.19	-1.31	1.69	-1.75	2.19	-1.75	2.25	-1.75	2.25	-1.75	2.25	-1.75	2.25	-1.75	2.25	-1.75	2.25
2008	2.5	0.19	0.19	-0.31	0.69	-0.81	1.19	-1.31	1.69	-1.81	2.19	-2.25	2.69	-2.25	2.75	-2.25	2.75	-2.25	2.75	-2.25	2.75	-2.25	2.75
2008	3.0	0.19	0.19	-0.31	0.69	-0.81	1.19	-1.31	1.69	-1.81	2.19	-2.31	2.69	-2.75	3.19	-2.75	3.25	-2.75	3.25	-2.75	3.25	-2.75	3.25
2008	3.5	0.19	0.19	-0.31	0.69	-0.81	1.19	-1.31	1.69	-1.81	2.19	-2.31	2.69	-2.81	3.19	-3.25	3.69	-3.25	3.75	-3.25	3.75	-3.25	3.75
2008	4.0	0.19	0.19	-0.31	0.69	-0.81	1.19	-1.31	1.69	-1.81	2.19	-2.31	2.69	-2.81	3.19	-3.31	3.69	-3.75	4.19	-3.75	4.25	-3.75	4.25
2008	4.5	0.19	0.19	-0.31	0.69	-0.81	1.19	-1.31	1.69	-1.81	2.19	-2.31	2.69	-2.81	3.19	-3.31	3.69	-3.81	4.19	-4.25	4.69	-4.25	4.75
2008	5.0	0.19	0.19	-0.31	0.69	-0.81	1.19	-1.31	1.69	-1.81	2.19	-2.31	2.69	-2.81	3.19	-3.31	3.69	-3.81	4.19	-4.31	4.69	-4.75	5.19
2009	0.0	1.38	0.77	0.88	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77
2009	0.5	1.38	1.27	0.88	1.27	0.38	1.27	0.27	1.27	0.27	1.27	0.27	1.27	0.27	1.27	0.27	1.27	0.27	1.27	0.27	1.27	0.27	1.27
2009	1.0	1.38	1.38	0.88	1.77	0.38	1.77	-0.12	1.77	-0.23	1.77	-0.23	1.77	-0.23	1.77	-0.23	1.77	-0.23	1.77	-0.23	1.77	-0.23	1.77
2009	1.5	1.38	1.38	0.88	1.88	0.38	2.27	-0.12	2.27	-0.62	2.27	-0.73	2.27	-0.73	2.27	-0.73	2.27	-0.73	2.27	-0.73	2.27	-0.73	2.27
2009	2.0	1.38	1.38	0.88	1.88	0.38	2.38	-0.12	2.77	-0.62	2.77	-1.12	2.77	-1.23	2.77	-1.23	2.77	-1.23	2.77	-1.23	2.77	-1.23	2.77
2009	2.5	1.38	1.38	0.88	1.88	0.38	2.38	-0.12	2.88	-0.62	3.27	-1.12	3.27	-1.62	3.27	-1.73	3.27	-1.73	3.27	-1.73	3.27	-1.73	3.27
2009	3.0	1.38	1.38	0.88	1.88	0.38	2.38	-0.12	2.88	-0.62	3.38	-1.12	3.77	-1.62	3.77	-2.12	3.77	-2.23	3.77	-2.23	3.77	-2.23	3.77
2009	3.5	1.38	1.38	0.88	1.88	0.38	2.38	-0.12	2.88	-0.62	3.38	-1.12	3.88	-1.62	4.27	-2.12	4.27	-2.62	4.27	-2.73	4.27	-2.73	4.27
2009	4.0	1.38	1.38	0.88	1.88	0.38	2.38	-0.12	2.88	-0.62	3.38	-1.12	3.88	-1.62	4.38	-2.12	4.77	-2.62	4.77	-3.12	4.77	-3.23	4.77

(Continues)

Table F.4 (Continued)

Year	δ_{SC}	0		0.5		1		1.5		2		2.5		3		3.5		4		4.5		5	
	δ_T	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2009	4.5	1.38	1.38	0.88	1.88	0.38	2.38	-0.12	2.88	-0.62	3.38	-1.12	3.88	-1.62	4.38	-2.12	4.88	-2.62	5.27	-3.12	5.27	-3.62	5.27
2009	5.0	1.38	1.38	0.88	1.88	0.38	2.38	-0.12	2.88	-0.62	3.38	-1.12	3.88	-1.62	4.38	-2.12	4.88	-2.62	5.38	-3.12	5.77	-3.62	5.77
2010	0.0	1.00	0.46	0.50	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46
2010	0.5	1.00	0.96	0.50	0.96	-0.00	0.96	-0.04	0.96	-0.04	0.96	-0.04	0.96	-0.04	0.96	-0.04	0.96	-0.04	0.96	-0.04	0.96	-0.04	0.96
2010	1.0	1.00	1.00	0.50	1.46	-0.00	1.46	-0.50	1.46	-0.54	1.46	-0.54	1.46	-0.54	1.46	-0.54	1.46	-0.54	1.46	-0.54	1.46	-0.54	1.46
2010	1.5	1.00	1.00	0.50	1.50	-0.00	1.96	-0.50	1.96	-1.00	1.96	-1.04	1.96	-1.04	1.96	-1.04	1.96	-1.04	1.96	-1.04	1.96	-1.04	1.96
2010	2.0	1.00	1.00	0.50	1.50	-0.00	2.00	-0.50	2.46	-1.00	2.46	-1.50	2.46	-1.54	2.46	-1.54	2.46	-1.54	2.46	-1.54	2.46	-1.54	2.46
2010	2.5	1.00	1.00	0.50	1.50	-0.00	2.00	-0.50	2.50	-1.00	2.96	-1.50	2.96	-2.00	2.96	-2.04	2.96	-2.04	2.96	-2.04	2.96	-2.04	2.96
2010	3.0	1.00	1.00	0.50	1.50	-0.00	2.00	-0.50	2.50	-1.00	3.00	-1.50	3.46	-2.00	3.46	-2.50	3.46	-2.54	3.46	-2.54	3.46	-2.54	3.46
2010	3.5	1.00	1.00	0.50	1.50	-0.00	2.00	-0.50	2.50	-1.00	3.00	-1.50	3.50	-2.00	3.96	-2.50	3.96	-3.00	3.96	-3.04	3.96	-3.04	3.96
2010	4.0	1.00	1.00	0.50	1.50	-0.00	2.00	-0.50	2.50	-1.00	3.00	-1.50	3.50	-2.00	4.00	-2.50	4.46	-3.00	4.46	-3.50	4.46	-3.54	4.46
2010	4.5	1.00	1.00	0.50	1.50	-0.00	2.00	-0.50	2.50	-1.00	3.00	-1.50	3.50	-2.00	4.00	-2.50	4.50	-3.00	4.96	-3.50	4.96	-4.00	4.96
2010	5.0	1.00	1.00	0.50	1.50	-0.00	2.00	-0.50	2.50	-1.00	3.00	-1.50	3.50	-2.00	4.00	-2.50	4.50	-3.00	5.00	-3.50	5.46	-4.00	5.46
2011	0.0	1.56	1.52	1.56	1.56	1.56	1.56	1.56	1.56	1.56	1.56	1.56	1.56	1.56	1.56	1.56	1.56	1.56	1.56	1.56	1.56	1.56	1.56
2011	0.5	1.52	1.52	1.06	2.02	1.06	2.06	1.06	2.06	1.06	2.06	1.06	2.06	1.06	2.06	1.06	2.06	1.06	2.06	1.06	2.06	1.06	2.06
2011	1.0	1.52	1.52	1.02	2.02	0.56	2.52	0.56	2.56	0.56	2.56	0.56	2.56	0.56	2.56	0.56	2.56	0.56	2.56	0.56	2.56	0.56	2.56
2011	1.5	1.52	1.52	1.02	2.02	0.52	2.52	0.06	3.02	0.06	3.06	0.06	3.06	0.06	3.06	0.06	3.06	0.06	3.06	0.06	3.06	0.06	3.06
2011	2.0	1.52	1.52	1.02	2.02	0.52	2.52	0.02	3.02	-0.44	3.52	-0.44	3.56	-0.44	3.56	-0.44	3.56	-0.44	3.56	-0.44	3.56	-0.44	3.56
2011	2.5	1.52	1.52	1.02	2.02	0.52	2.52	0.02	3.02	-0.48	3.52	-0.94	4.02	-0.94	4.06	-0.94	4.06	-0.94	4.06	-0.94	4.06	-0.94	4.06
2011	3.0	1.52	1.52	1.02	2.02	0.52	2.52	0.02	3.02	-0.48	3.52	-0.98	4.02	-1.44	4.52	-1.44	4.56	-1.44	4.56	-1.44	4.56	-1.44	4.56
2011	3.5	1.52	1.52	1.02	2.02	0.52	2.52	0.02	3.02	-0.48	3.52	-0.98	4.02	-1.48	4.52	-1.94	5.02	-1.94	5.06	-1.94	5.06	-1.94	5.06
2011	4.0	1.52	1.52	1.02	2.02	0.52	2.52	0.02	3.02	-0.48	3.52	-0.98	4.02	-1.48	4.52	-1.98	5.02	-2.44	5.52	-2.44	5.56	-2.44	5.56
2011	4.5	1.52	1.52	1.02	2.02	0.52	2.52	0.02	3.02	-0.48	3.52	-0.98	4.02	-1.48	4.52	-1.98	5.02	-2.48	5.52	-2.94	6.02	-2.94	6.06
2011	5.0	1.52	1.52	1.02	2.02	0.52	2.52	0.02	3.02	-0.48	3.52	-0.98	4.02	-1.48	4.52	-1.98	5.02	-2.48	5.52	-2.98	6.02	-3.44	6.52
2012	0.0	2.24	1.49	2.24	1.99	2.24	2.24	2.24	2.24	2.24	2.24	2.24	2.24	2.24	2.24	2.24	2.24	2.24	2.24	2.24	2.24	2.24	2.24
2012	0.5	1.74	1.49	1.74	1.99	1.74	2.49	1.74	2.74	1.74	2.74	1.74	2.74	1.74	2.74	1.74	2.74	1.74	2.74	1.74	2.74	1.74	2.74
2012	1.0	1.49	1.49	1.24	1.99	1.24	2.49	1.24	2.99	1.24	3.24	1.24	3.24	1.24	3.24	1.24	3.24	1.24	3.24	1.24	3.24	1.24	3.24
2012	1.5	1.49	1.49	0.99	1.99	0.74	2.49	0.74	2.99	0.74	3.49	0.74	3.74	0.74	3.74	0.74	3.74	0.74	3.74	0.74	3.74	0.74	3.74
2012	2.0	1.49	1.49	0.99	1.99	0.49	2.49	0.24	2.99	0.24	3.49	0.24	3.99	0.24	4.24	0.24	4.24	0.24	4.24	0.24	4.24	0.24	4.24
2012	2.5	1.49	1.49	0.99	1.99	0.49	2.49	-0.01	2.99	-0.26	3.49	-0.26	3.99	-0.26	4.49	-0.26	4.74	-0.26	4.74	-0.26	4.74	-0.26	4.74
2012	3.0	1.49	1.49	0.99	1.99	0.49	2.49	-0.01	2.99	-0.51	3.49	-0.76	3.99	-0.76	4.49	-0.76	4.99	-0.76	5.24	-0.76	5.24	-0.76	5.24

(Continues)

Table F.4 (Continued)

Year	δ_{SC}	0		0.5		1		1.5		2		2.5		3		3.5		4		4.5		5	
	δ_T	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2012	3.5	1.49	1.49	0.99	1.99	0.49	2.49	-0.01	2.99	-0.51	3.49	-1.01	3.99	-1.26	4.49	-1.26	4.99	-1.26	5.49	-1.26	5.74	-1.26	5.74
2012	4.0	1.49	1.49	0.99	1.99	0.49	2.49	-0.01	2.99	-0.51	3.49	-1.01	3.99	-1.51	4.49	-1.76	4.99	-1.76	5.49	-1.76	5.99	-1.76	6.24
2012	4.5	1.49	1.49	0.99	1.99	0.49	2.49	-0.01	2.99	-0.51	3.49	-1.01	3.99	-1.51	4.49	-2.01	4.99	-2.26	5.49	-2.26	5.99	-2.26	6.49
2012	5.0	1.49	1.49	0.99	1.99	0.49	2.49	-0.01	2.99	-0.51	3.49	-1.01	3.99	-1.51	4.49	-2.01	4.99	-2.51	5.49	-2.76	5.99	-2.76	6.49
2013	0.0	1.99	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89
2013	0.5	1.99	1.99	1.49	2.39	1.39	2.39	1.39	2.39	1.39	2.39	1.39	2.39	1.39	2.39	1.39	2.39	1.39	2.39	1.39	2.39	1.39	2.39
2013	1.0	1.99	1.99	1.49	2.49	0.99	2.89	0.89	2.89	0.89	2.89	0.89	2.89	0.89	2.89	0.89	2.89	0.89	2.89	0.89	2.89	0.89	2.89
2013	1.5	1.99	1.99	1.49	2.49	0.99	2.99	0.49	3.39	0.39	3.39	0.39	3.39	0.39	3.39	0.39	3.39	0.39	3.39	0.39	3.39	0.39	3.39
2013	2.0	1.99	1.99	1.49	2.49	0.99	2.99	0.49	3.49	-0.01	3.89	-0.11	3.89	-0.11	3.89	-0.11	3.89	-0.11	3.89	-0.11	3.89	-0.11	3.89
2013	2.5	1.99	1.99	1.49	2.49	0.99	2.99	0.49	3.49	-0.01	3.99	-0.51	4.39	-0.61	4.39	-0.61	4.39	-0.61	4.39	-0.61	4.39	-0.61	4.39
2013	3.0	1.99	1.99	1.49	2.49	0.99	2.99	0.49	3.49	-0.01	3.99	-0.51	4.49	-1.01	4.89	-1.11	4.89	-1.11	4.89	-1.11	4.89	-1.11	4.89
2013	3.5	1.99	1.99	1.49	2.49	0.99	2.99	0.49	3.49	-0.01	3.99	-0.51	4.49	-1.01	4.99	-1.51	5.39	-1.61	5.39	-1.61	5.39	-1.61	5.39
2013	4.0	1.99	1.99	1.49	2.49	0.99	2.99	0.49	3.49	-0.01	3.99	-0.51	4.49	-1.01	4.99	-1.51	5.49	-2.01	5.89	-2.11	5.89	-2.11	5.89
2013	4.5	1.99	1.99	1.49	2.49	0.99	2.99	0.49	3.49	-0.01	3.99	-0.51	4.49	-1.01	4.99	-1.51	5.49	-2.01	5.99	-2.51	6.39	-2.61	6.39
2013	5.0	1.99	1.99	1.49	2.49	0.99	2.99	0.49	3.49	-0.01	3.99	-0.51	4.49	-1.01	4.99	-1.51	5.49	-2.01	5.99	-2.51	6.49	-3.01	6.89
2014	0.0	2.43	1.44	1.93	1.44	1.44	1.44	1.44	1.44	1.44	1.44	1.44	1.44	1.44	1.44	1.44	1.44	1.44	1.44	1.44	1.44	1.44	1.44
2014	0.5	2.43	1.94	1.93	1.94	1.43	1.94	0.94	1.94	0.94	1.94	0.94	1.94	0.94	1.94	0.94	1.94	0.94	1.94	0.94	1.94	0.94	1.94
2014	1.0	2.43	2.43	1.93	2.44	1.43	2.44	0.93	2.44	0.44	2.44	0.44	2.44	0.44	2.44	0.44	2.44	0.44	2.44	0.44	2.44	0.44	2.44
2014	1.5	2.43	2.43	1.93	2.93	1.43	2.94	0.93	2.94	0.43	2.94	-0.06	2.94	-0.06	2.94	-0.06	2.94	-0.06	2.94	-0.06	2.94	-0.06	2.94
2014	2.0	2.43	2.43	1.93	2.93	1.43	3.43	0.93	3.44	0.43	3.44	-0.07	3.44	-0.56	3.44	-0.56	3.44	-0.56	3.44	-0.56	3.44	-0.56	3.44
2014	2.5	2.43	2.43	1.93	2.93	1.43	3.43	0.93	3.93	0.43	3.94	-0.07	3.94	-0.57	3.94	-1.06	3.94	-1.06	3.94	-1.06	3.94	-1.06	3.94
2014	3.0	2.43	2.43	1.93	2.93	1.43	3.43	0.93	3.93	0.43	4.43	-0.07	4.44	-0.57	4.44	-1.07	4.44	-1.56	4.44	-1.56	4.44	-1.56	4.44
2014	3.5	2.43	2.43	1.93	2.93	1.43	3.43	0.93	3.93	0.43	4.43	-0.07	4.93	-0.57	4.94	-1.07	4.94	-1.57	4.94	-2.06	4.94	-2.06	4.94
2014	4.0	2.43	2.43	1.93	2.93	1.43	3.43	0.93	3.93	0.43	4.43	-0.07	4.93	-0.57	5.43	-1.07	5.44	-1.57	5.44	-2.07	5.44	-2.56	5.44
2014	4.5	2.43	2.43	1.93	2.93	1.43	3.43	0.93	3.93	0.43	4.43	-0.07	4.93	-0.57	5.43	-1.07	5.93	-1.57	5.94	-2.07	5.94	-2.57	5.94
2014	5.0	2.43	2.43	1.93	2.93	1.43	3.43	0.93	3.93	0.43	4.43	-0.07	4.93	-0.57	5.43	-1.07	5.93	-1.57	6.43	-2.07	6.44	-2.57	6.44
2015	0.0	3.00	2.60	2.60	2.60	2.60	2.60	2.60	2.60	2.60	2.60	2.60	2.60	2.60	2.60	2.60	2.60	2.60	2.60	2.60	2.60	2.60	2.60
2015	0.5	3.00	3.00	2.50	3.10	2.10	3.10	2.10	3.10	2.10	3.10	2.10	3.10	2.10	3.10	2.10	3.10	2.10	3.10	2.10	3.10	2.10	3.10
2015	1.0	3.00	3.00	2.50	3.50	2.00	3.60	1.60	3.60	1.60	3.60	1.60	3.60	1.60	3.60	1.60	3.60	1.60	3.60	1.60	3.60	1.60	3.60
2015	1.5	3.00	3.00	2.50	3.50	2.00	4.00	1.50	4.10	1.10	4.10	1.10	4.10	1.10	4.10	1.10	4.10	1.10	4.10	1.10	4.10	1.10	4.10
2015	2.0	3.00	3.00	2.50	3.50	2.00	4.00	1.50	4.50	1.00	4.60	0.60	4.60	0.60	4.60	0.60	4.60	0.60	4.60	0.60	4.60	0.60	4.60

(Continues)

Table F.4 (Continued)

Year	δ_{SC}	0		0.5		1		1.5		2		2.5		3		3.5		4		4.5		5	
	δ_T	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2015	2.5	3.00	3.00	2.50	3.50	2.00	4.00	1.50	4.50	1.00	5.00	0.50	5.10	0.10	5.10	0.10	5.10	0.10	5.10	0.10	5.10	0.10	5.10
2015	3.0	3.00	3.00	2.50	3.50	2.00	4.00	1.50	4.50	1.00	5.00	0.50	5.50	0.00	5.60	-0.40	5.60	-0.40	5.60	-0.40	5.60	-0.40	5.60
2015	3.5	3.00	3.00	2.50	3.50	2.00	4.00	1.50	4.50	1.00	5.00	0.50	5.50	0.00	6.00	-0.50	6.10	-0.90	6.10	-0.90	6.10	-0.90	6.10
2015	4.0	3.00	3.00	2.50	3.50	2.00	4.00	1.50	4.50	1.00	5.00	0.50	5.50	0.00	6.00	-0.50	6.50	-1.00	6.60	-1.40	6.60	-1.40	6.60
2015	4.5	3.00	3.00	2.50	3.50	2.00	4.00	1.50	4.50	1.00	5.00	0.50	5.50	0.00	6.00	-0.50	6.50	-1.00	7.00	-1.50	7.10	-1.90	7.10
2015	5.0	3.00	3.00	2.50	3.50	2.00	4.00	1.50	4.50	1.00	5.00	0.50	5.50	0.00	6.00	-0.50	6.50	-1.00	7.00	-1.50	7.50	-2.00	7.60
2016	0.0	2.46	1.55	1.96	1.55	1.55	1.55	1.55	1.55	1.55	1.55	1.55	1.55	1.55	1.55	1.55	1.55	1.55	1.55	1.55	1.55	1.55	1.55
2016	0.5	2.46	2.05	1.96	2.05	1.46	2.05	1.05	2.05	1.05	2.05	1.05	2.05	1.05	2.05	1.05	2.05	1.05	2.05	1.05	2.05	1.05	2.05
2016	1.0	2.46	2.46	1.96	2.55	1.46	2.55	0.96	2.55	0.55	2.55	0.55	2.55	0.55	2.55	0.55	2.55	0.55	2.55	0.55	2.55	0.55	2.55
2016	1.5	2.46	2.46	1.96	2.96	1.46	3.05	0.96	3.05	0.46	3.05	0.05	3.05	0.05	3.05	0.05	3.05	0.05	3.05	0.05	3.05	0.05	3.05
2016	2.0	2.46	2.46	1.96	2.96	1.46	3.46	0.96	3.55	0.46	3.55	-0.04	3.55	-0.45	3.55	-0.45	3.55	-0.45	3.55	-0.45	3.55	-0.45	3.55
2016	2.5	2.46	2.46	1.96	2.96	1.46	3.46	0.96	3.96	0.46	4.05	-0.04	4.05	-0.54	4.05	-0.95	4.05	-0.95	4.05	-0.95	4.05	-0.95	4.05
2016	3.0	2.46	2.46	1.96	2.96	1.46	3.46	0.96	3.96	0.46	4.46	-0.04	4.55	-0.54	4.55	-1.04	4.55	-1.45	4.55	-1.45	4.55	-1.45	4.55
2016	3.5	2.46	2.46	1.96	2.96	1.46	3.46	0.96	3.96	0.46	4.46	-0.04	4.96	-0.54	5.05	-1.04	5.05	-1.54	5.05	-1.95	5.05	-1.95	5.05
2016	4.0	2.46	2.46	1.96	2.96	1.46	3.46	0.96	3.96	0.46	4.46	-0.04	4.96	-0.54	5.46	-1.04	5.55	-1.54	5.55	-2.04	5.55	-2.45	5.55
2016	4.5	2.46	2.46	1.96	2.96	1.46	3.46	0.96	3.96	0.46	4.46	-0.04	4.96	-0.54	5.46	-1.04	5.96	-1.54	6.05	-2.04	6.05	-2.54	6.05
2016	5.0	2.46	2.46	1.96	2.96	1.46	3.46	0.96	3.96	0.46	4.46	-0.04	4.96	-0.54	5.46	-1.04	5.96	-1.54	6.46	-2.04	6.55	-2.54	6.55
2017	0.0	2.41	1.57	1.91	1.57	1.57	1.57	1.57	1.57	1.57	1.57	1.57	1.57	1.57	1.57	1.57	1.57	1.57	1.57	1.57	1.57	1.57	1.57
2017	0.5	2.41	2.07	1.91	2.07	1.41	2.07	1.07	2.07	1.07	2.07	1.07	2.07	1.07	2.07	1.07	2.07	1.07	2.07	1.07	2.07	1.07	2.07
2017	1.0	2.41	2.41	1.91	2.57	1.41	2.57	0.91	2.57	0.57	2.57	0.57	2.57	0.57	2.57	0.57	2.57	0.57	2.57	0.57	2.57	0.57	2.57
2017	1.5	2.41	2.41	1.91	2.91	1.41	3.07	0.91	3.07	0.41	3.07	0.07	3.07	0.07	3.07	0.07	3.07	0.07	3.07	0.07	3.07	0.07	3.07
2017	2.0	2.41	2.41	1.91	2.91	1.41	3.41	0.91	3.57	0.41	3.57	-0.09	3.57	-0.43	3.57	-0.43	3.57	-0.43	3.57	-0.43	3.57	-0.43	3.57
2017	2.5	2.41	2.41	1.91	2.91	1.41	3.41	0.91	3.91	0.41	4.07	-0.09	4.07	-0.59	4.07	-0.93	4.07	-0.93	4.07	-0.93	4.07	-0.93	4.07
2017	3.0	2.41	2.41	1.91	2.91	1.41	3.41	0.91	3.91	0.41	4.41	-0.09	4.57	-0.59	4.57	-1.09	4.57	-1.43	4.57	-1.43	4.57	-1.43	4.57
2017	3.5	2.41	2.41	1.91	2.91	1.41	3.41	0.91	3.91	0.41	4.41	-0.09	4.91	-0.59	5.07	-1.09	5.07	-1.59	5.07	-1.93	5.07	-1.93	5.07
2017	4.0	2.41	2.41	1.91	2.91	1.41	3.41	0.91	3.91	0.41	4.41	-0.09	4.91	-0.59	5.41	-1.09	5.57	-1.59	5.57	-2.09	5.57	-2.43	5.57
2017	4.5	2.41	2.41	1.91	2.91	1.41	3.41	0.91	3.91	0.41	4.41	-0.09	4.91	-0.59	5.41	-1.09	5.91	-1.59	6.07	-2.09	6.07	-2.59	6.07
2017	5.0	2.41	2.41	1.91	2.91	1.41	3.41	0.91	3.91	0.41	4.41	-0.09	4.91	-0.59	5.41	-1.09	5.91	-1.59	6.41	-2.09	6.57	-2.59	6.57
2018	0.0	2.47	0.19	1.97	0.19	1.47	0.19	0.97	0.19	0.47	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19
2018	0.5	2.47	0.69	1.97	0.69	1.47	0.69	0.97	0.69	0.47	0.69	-0.03	0.69	-0.31	0.69	-0.31	0.69	-0.31	0.69	-0.31	0.69	-0.31	0.69
2018	1.0	2.47	1.19	1.97	1.19	1.47	1.19	0.97	1.19	0.47	1.19	-0.03	1.19	-0.53	1.19	-0.81	1.19	-0.81	1.19	-0.81	1.19	-0.81	1.19

(Continues)

Table F.4 (Continued)

Year	δ_{SC}	0		0.5		1		1.5		2		2.5		3		3.5		4		4.5		5	
	δ_T	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
2018	1.5	2.47	1.69	1.97	1.69	1.47	1.69	0.97	1.69	0.47	1.69	-0.03	1.69	-0.53	1.69	-1.03	1.69	-1.31	1.69	-1.31	1.69	-1.31	1.69
2018	2.0	2.47	2.19	1.97	2.19	1.47	2.19	0.97	2.19	0.47	2.19	-0.03	2.19	-0.53	2.19	-1.03	2.19	-1.53	2.19	-1.81	2.19	-1.81	2.19
2018	2.5	2.47	2.47	1.97	2.69	1.47	2.69	0.97	2.69	0.47	2.69	-0.03	2.69	-0.53	2.69	-1.03	2.69	-1.53	2.69	-2.03	2.69	-2.31	2.69
2018	3.0	2.47	2.47	1.97	2.97	1.47	3.19	0.97	3.19	0.47	3.19	-0.03	3.19	-0.53	3.19	-1.03	3.19	-1.53	3.19	-2.03	3.19	-2.53	3.19
2018	3.5	2.47	2.47	1.97	2.97	1.47	3.47	0.97	3.69	0.47	3.69	-0.03	3.69	-0.53	3.69	-1.03	3.69	-1.53	3.69	-2.03	3.69	-2.53	3.69
2018	4.0	2.47	2.47	1.97	2.97	1.47	3.47	0.97	3.97	0.47	4.19	-0.03	4.19	-0.53	4.19	-1.03	4.19	-1.53	4.19	-2.03	4.19	-2.53	4.19
2018	4.5	2.47	2.47	1.97	2.97	1.47	3.47	0.97	3.97	0.47	4.47	-0.03	4.69	-0.53	4.69	-1.03	4.69	-1.53	4.69	-2.03	4.69	-2.53	4.69
2018	5.0	2.47	2.47	1.97	2.97	1.47	3.47	0.97	3.97	0.47	4.47	-0.03	4.97	-0.53	5.19	-1.03	5.19	-1.53	5.19	-2.03	5.19	-2.53	5.19

Distributional Difference-in-Differences with Multiple Time Periods: A Monte Carlo Analysis

Andrea Ciaccio[1]

[1]Department of Economics, Ca' Foscari University of Venice

Abstract

Researchers are often interested in evaluating the impact of a policy on the entire (or specific parts of the) distribution of the outcome of interest. In this paper, I provide a method to recover the whole distribution of the untreated potential outcome for the treated group in non-experimental settings with staggered treatment adoption by generalizing the existing quantile treatment effects on the treated (QTT) estimator proposed by Callaway and Li (2019). Besides the QTT, I consider different approaches that anonymously summarize the quantiles of the distribution of the outcome of interest (such as tests for stochastic dominance rankings) without relying on rank invariance assumptions. The finite-sample properties of the estimator proposed are analyzed via different Monte Carlo simulations. Despite being slightly biased for relatively small sample sizes, the proposed method's performance increases substantially when the sample size increases.

JEL codes: C14, C21, C23

Keywords: Quantile treatment effect on the treated, Difference in Differences, Copula, Variation in treatment timing, Treatment effect heterogeneity, Rank Invariance

Acknowledgments: I am grateful to Irene Mammi, Davide Raggi, Federico Belotti, Paolo Li Donni, Mario Padula, Manuel Arellano, and Dmitry Arkhangelsky for their invaluable advice, insights, constructive suggestions, and detailed feedback on previous drafts of this paper. I also thank Alexander Simon Mayer, Annalivia Polselli, and Francesco Chiocchio for their thoughtful comments on an earlier version. Additionally, I appreciate the helpful feedback from participants at the CEMFI Econometrics Workshop and the Ca' Foscari Internal Departmental Seminar.

2.1 Introduction

In economics and, more generally, in social sciences, we are often interested in assessing the impact of a policy intervention on an outcome of interest in non-experimental settings. However, researchers face a selection problem when assessing the policy's causal effect. Once the policy is implemented, data will not reveal counterfactual outcomes (i.e., the outcomes we would have observed had the policy not been implemented) (Holland, 1986). The researcher's main challenge will be finding an appropriate way to recover the missing counterfactual outcomes.

Different methods exist to estimate the causal effect of policy interventions in non-experimental settings. One of the most popular methods is the *Difference-in-Differences (DiD)*. In the classical DiD setting, we have two periods and two groups: in the first period, no unit is treated; in the second period, some units become treated (treated group) while others remain untreated (untreated group). Under the assumption that had the policy not been implemented, the average evolution between treated and untreated units would have been parallel (also known as *common trends* or *parallel trends (PT)* assumption), it is possible to estimate the counterfactual outcome for the treated group consistently. Once the counterfactual outcome is retrieved, the most common treatment effect parameter considered in these settings is the Average Treatment Effect on the Treated (ATT), which estimates the average causal effect of the policy for the treated subpopulation.[1] Specifically, estimation of the ATT is achieved by comparing the average change in the outcomes observed for the treated subpopulation to that experienced by the untreated group.

While most studies employ causal inference methods to estimate the ATT, sometimes researchers might also be interested in evaluating the impact of a policy on the entire (or specific parts of the) distribution of the outcome of interest. For instance, think of two policies that aim at increasing wages but with the same average impact. Policymakers will prefer the policy that is more likely to lead to a more significant increase in the lower deciles of the income distribution than the one that should generate higher (expected) benefits for individuals "dwelling" in the middle-top deciles. Similarly, there are plenty of situations in economics where considering the entire distributional effect of a policy is more appropriate, especially when there are reasons to expect the impact to be heterogeneous across treated groups.

One way to examine the distributional effect of a policy is to consider the Quantile Treatment Effect on the Treated (QTT). The QTT estimates the causal effect of a policy – for the treated group – on a specific quantile of the outcome of interest by comparing the quantiles of treated and untreated outcomes. Different studies exist that show how to retrieve a consistent estimator of the QTT in two-period two-group settings (Athey and Imbens, 2006; Bonhomme and Sauder, 2011; Fan and Yu, 2012; Callaway et al., 2018; Callaway and Li, 2019).

In this paper, I propose a method to recover counterfactual quantiles and, more generally, the entire distribution of the counterfactual outcome for the treated group in non-experimental settings with multiple groups and periods, where treatment timing varies. Many non-experimental designs in empirical research deviate from the canonical (2×2) scenario. A strand of the literature on causal inference has recently focused on how departures from the (2×2) scenario influence the estimation of the

[1] The DiD's popularity is mainly due to its broad applicability to many research questions and the fact that, if we have many independent clusters, one can prove that we can consistently estimate the ATT using a two-way fixed effects (TWFE) regression.

ATT (Borusyak et al., 2021; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Wooldridge, 2021; De Chaisemartin and d'Haultfoeuille, 2022).[2]

While several papers provide methods for obtaining a consistent estimator of the ATT in staggered policy rollout contexts, little is known about how to estimate the distributional effect in these settings. The only study I am aware of that addresses the identification of the QTT in staggered DiD settings is Li and Lin (2024). Although their identification result is similar to the one presented in this paper, they do not offer an estimator for the QTT, limiting the practical applicability of their findings. A detailed comparison with their work is provided at the end of this section, following the presentation of the identification results in this paper.

To identify and estimate the full counterfactual distribution in settings with staggered treatment adoption, I exploit the intuition behind the estimator of the group-time average treatment effects proposed by Callaway and Sant'Anna (2021). By applying to each pair of treated cohorts and never-treated units the method proposed by Callaway and Li (2019), I show that identification and estimation of the entire counterfactual distribution can be achieved.

Specifically, Callaway and Li (2019) achieve identification and estimation of the counterfactual distribution under a distributional version of the PT trends assumption. The authors require the change in the untreated potential outcome to be independent of treatment assignment. The similarity of the distributional PT assumption with the canonical (mean) PT assumption makes this method very intuitive.

Compared to the selection on observables and strong ignorability assumptions invoked by Firpo (2007), the distributional parallel trends allow unobservable confounders to vary between treated and untreated units. For instance, in the context analyzed by Bonhomme and Sauder (2011) of how selective and non-selective secondary education impact children's test scores, selection on unobservables means that a child's unobserved initial endowment can be (potentially) correlated with the type of education in which the pupil enrolls.

Despite the distributional PT assumption being more than sufficient to reach point identification of the ATT, Fan and Yu (2012) show that the distributional treatment effect on the treated is only partially identified. Partial identification arises from the fact that the dependence (or copula) between the change in untreated potential outcome and the pre-treatment level of untreated potential outcome is unknown to the researcher. To reach point identification, Callaway and Li (2019) impose a copula stability assumption, which requires this (missing) dependence to be stable over time. Under the availability of (at least) two pre-treatment periods, the researcher can recover the missing dependence using the (known) dependence in the previous periods.

One drawback of this approach, however, is that it requires access to panel data and at least two pre-treatment periods. To overcome this requirement, following Callaway et al. (2018), I assume *copula invariance*, which implies that the missing dependence structure is the same for treated and never-treated units.

For instance, suppose that among never-treated units, the largest earnings increases over time occur for those in the lower deciles of the pre-treatment income distribution. The copula invariance assumption requires that, in the absence of the policy, the same pattern would have emerged for treated units.

Compared to assuming copula stability, adopting copula invariance has the additional advantage of

[2]See, for instance, Roth et al. (2023) for a review.

extending the analysis to settings where repeated cross-sections, rather than panel data, are available. Under the additional assumption of rank invariance of potential outcomes over time, I demonstrate that the main identification results of this paper can indeed be extended to cases where only repeated cross-sections are available.

It is important to emphasize that the copula invariance assumption does not restrict the marginal distributions; it solely requires that the dependence structure remain the same across treated and never-treated units.

Furthermore, similarly to Callaway and Li (2019), I extend the results obtained to the case where the distributional PT and the copula invariance assumption may hold after conditioning on observable characteristics.

Once the entire counterfactual distribution is identified, the researcher can construct different causal parameters of interest. For instance, a generalization of the QTT estimator proposed by Callaway and Li (2019) to the multi-periods and multiple groups set-up can be constructed. I call this parameter *cohort-time quantile treatment effect* in the spirit of the Callaway and Sant'Anna (2021)'s causal parameters.

When computing the QTT, however, the researcher implicitly invokes an additional assumption: rank invariance between treated and untreated groups. As all these studies exploit information observed in the untreated group to infer the counterfactual distribution for the treated group, the rank invariance assumption states that treated and untreated units occupying the same ranking are comparable. However, the fact that treated and untreated units have the same ranking does not necessarily mean that rank is preserved when, for instance, treated units are endowed with untreated units' abilities. As Maasoumi and Wang (2019) show, the data often reject rank invariance and, therefore, is hard to justify in most empirical applications.

For this reason, in this paper, building on Maasoumi and Wang (2019), I also consider alternative causal parameters that do not rely on the assumption of rank invariance. Specifically, I demonstrate that once the entire counterfactual distribution is estimated, researchers can construct causal estimands based on inequality measures – such as the Lorenz curve, Gini coefficient, and others. For instance, one can define a causal estimand by contrasting the Gini indices computed from the CDFs of the treated and untreated potential outcomes for the treated group.

Inequality measures characterize the quantiles of the distribution of interest *anonymously*–that is, without specifying the identity of units occupying a particular quantile. This approach addresses the limitation of rank invariance by focusing on the distributions of treated and untreated potential outcomes rather than on specific units.

However, since there is no universally accepted distribution evaluation function for conciseness, this paper will focus solely on *tests for stochastic dominance rankings* to assess the distributional treatment effect on the treated. To the best of my knowledge, this is the first study to integrate the literature on causal inference with that on inequality measures.

Depending on the causal parameter of interest, I advocate for the most appropriate statistical test to conduct valid inference. For instance, when considering the QTT, the empirical bootstrap proposed by Callaway and Li (2019) will be considered.

Once the causal parameters of interest are built, one can aggregate these parameters using aggregation schemes identical to those proposed by Callaway and Sant'Anna (2021) to highlight heterogeneity along specific dimensions (such as how the treatment effects vary with the length of exposure to the

treatment). This step is particularly crucial when there are many parameters to estimate, and the researcher wants to have a tool to aggregate and summarize results to highlight heterogeneity along a given dimension (e.g., over time) and/or to assess the overall impact of the policy.

Based on the identification results of this paper, estimating the entire counterfactual distribution is simple as it relies on non-parametric estimation of empirical distribution functions. When the distributional PT is likely to hold after conditioning on pre-treatment covariates, estimation of the counterfactual distribution is performed via a generalization of the Inverse Probability Weighting estimand first proposed by Firpo (2007) and then readapted by Callaway and Li (2019). When the copula invariance assumption also holds conditionally, one can exploit the intuition behind Callaway et al. (2018). Specifically, assuming that covariates are discrete, one can compute the conditional counterfactual distribution for each possible value of the covariates.

Via different Monte Carlo exercises, I show that, despite being slightly biased for relatively small sample sizes, the proposed method's performance increases substantially as the sample size increases. In addition, I show that this method's performance is also robust to small deviations from the main identifying assumptions.

Given the similarity between the identification results in this paper and those in the concurrent study by Li and Lin (2024), it is important to highlight the key differences. First, while I generalize Callaway and Li (2019), they extend the method proposed in Callaway et al. (2018). However, both approaches rely on a Distributional PT assumption and a Copula Invariance assumption for identification. The use of Copula Invariance in both papers enables extending the analysis to settings with repeated cross-sections.

A key distinction is that Li and Lin (2024) present identification results based on a not-yet-treated comparison group, whereas I focus on a never-treated comparison group. However, while I primarily present identification results for the setting with a never-treated comparison group, these results can be generalized to cases where not-yet-treated units serve as the comparison group, as they follow from symmetric arguments. Additionally, while using not-yet-treated units expands the set of valid comparison units–potentially leading to more informative inference due to their likely similarity in pre-treatment trends with already-treated units–this approach has limitations. If later-treated units adjust their behavior after the policy's introduction, the never-treated group remains a more suitable comparison.

Additionally, while both studies rely on Copula Invariance for QTT identification, I explicitly acknowledge the computational challenges when requiring the assumption to hold conditionally. As noted in Callaway and Li (2019), estimating the QTT under Conditional Copula Invariance requires estimating five conditional distributions, which is often infeasible in empirical applications. For this reason, following Callaway and Li (2019), I provide three different identification results depending on whether: (i) both the Distributional PT and Copula Invariance assumptions hold unconditionally, (ii) the Distributional PT assumption holds conditionally on covariates while Copula Invariance remains unconditional, or (iii) both assumptions hold conditionally on observed covariates. In the third scenario, I demonstrate how to address the computational challenge mentioned above.

While Callaway et al. (2018) allow for Copula Invariance to hold conditionally on covariates, their identification results focus on estimating a conditional QTT, whereas Li and Lin (2024) examine the unconditional QTT. Additionally, applying Callaway et al. (2018)'s method requires discrete covariates. Since Li and Lin (2024) do not propose an estimator for the QTT, as I do, nor provide any details

regarding the nature of covariates, it remains unclear how to estimate the unconditional QTT under Conditional Copula Invariance, limiting the practical applicability of their identification results.

Lastly, compared to this latter study, I propose aggregation schemes similar to those in Callaway and Sant'Anna (2021) to highlight heterogeneity along specific dimensions. I also emphasize that once the entire counterfactual distribution is identified, researchers can construct various causal parameters beyond the QTT, which typically relies on the strong assumption of rank invariance.

Thus, while the identification results in this paper are similar to those in Li and Lin (2024), this paper offers greater generality and broader applicability.

Because I focus on non-parametric estimation of distributional treatment effect in a context with multiple periods and multiple groups, this paper is linked to different strands of literature. First, as pointed out in Callaway and Li (2019), it contributes to the literature on nonseparable panel data methods (see, for instance, Chernozhukov et al., 2013). Second, it is related to two strands of the literature in causal inference: the one which focuses on the estimation of distributional treatment effects while still allowing for selection on unobservables (Athey and Imbens, 2006; Bonhomme and Sauder, 2011; Callaway et al., 2018; Callaway and Li, 2019); as well as the one on staggered treatment adoption (Borusyak et al., 2021; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Wooldridge, 2021; De Chaisemartin and d'Haultfoeuille, 2022). Lastly, since I also consider parameters that do not require rank invariance assumption to hold (such as tests of stochastic dominance), this paper contributes to the ongoing debate on inequality measures (see, for instance, Maasoumi and Wang, 2019).

This paper is organized as follows. Section 2.2 presents the main identification results and introduces different aggregation schemes to capture heterogeneity along specific dimensions. Section 2.3 outlines the estimation and inference procedures. Section 2.4 evaluates the finite-sample properties of the proposed estimators through various Monte Carlo exercises. In Section 2.5, I compare the proposed estimator with existing methods in the literature on distributional treatment effects, and the estimator proposed in Callaway and Sant'Anna (2021). Finally, Section 2.6 provides concluding remarks.

All proofs are presented in Appendix A, while Appendix B extends the analysis to settings with repeated cross-sections.

2.2 Identification

This section presents the main identification results of this paper. Specifically, I will provide the identification of the full distribution of the untreated potential outcome for the treated group, both when the identifying assumptions hold unconditionally and when they hold after conditioning on observed characteristics.

Throughout the paper, for any two random variables T and Z, I denote the support of T by $supp(T)$, and the support of T given Z by $supp(T|Z)$.

2.2.1 Setup

Let us consider the case where there are available T periods, and let us denote a generic period with $t = 1, \dots, T$. Let $D_{i,t}$ denote the treatment dummy taking value 1 if unit i is treated in period t, where $i = 1, \dots, N$. In the classic (2×2) DiD scenario, $T = 2$ and no units are treated in $t = 1$. However, in the context analyzed in the rest of this paper, I will allow $T > 2$. In particular, I will

assume that starting from period $q \geq 2$, there is a staggered policy rollout, where q denotes the first period the policy intervention is implemented. Then, one should make the following assumption about the treatment process:

Assumption 1 (*Irreversibility of Treatment*). *$D_j = 0$ for all $j = 1, \ldots, (q-1)$ almost surely (a.s.). For $t = q, \ldots, T$, $D_{t-1} = 1$ implies that $D_t = 1$ a.s.*

Assumption 1 states that no unit is treated before period q and that once a unit receives the treatment, it will remain treated for the remainder of the panel. This assumption is often referred to as *staggered treatment adoption*. Following Callaway and Sant'Anna (2021) and Sun and Abraham (2021), I will interpret this assumption as units changing their behavior "forever" once they become treated.[3]

Another implication of Assumption 1 is that it automatically defines the "cohort" to which unit i belongs. Assuming a never-treated group exists, one can define $T - q + 2$ mutually exclusive cohort dummies, d_r (with $r = q, \ldots, T$), denoting the period in which unit i first receives the treatment. C will, on the other hand, be a dummy taking value 1 if a group is never-treated. Lastly, $\bar{d} = \max_{i=1,\ldots,N} d_{i,r}$ denotes the maximum period in the dataset where a unit may become treated.

For brevity, I will present the main identification results focusing on the setting with a never-treated comparison group. These results can be generalized to cases where not-yet-treated units serve as the comparison group, as they follow from symmetric arguments, as noted by Callaway and Sant'Anna (2021). While using not-yet-treated units allows for a larger set of valid comparison units–potentially leading to more informative inference, given their likely similarity in pre-treatment trends with already-treated units–this approach has a limitation. If later treated units adjust their behavior after the policy's introduction, the never-treated group remains a more suitable comparison.[4]

Next, assume the existence of a complete set of pre-treatment covariates, denoted by X with support denoted by $\chi = supp(X) \subseteq R^K$ and dimensionality equal to $dim(X) = k$. The generalized propensity score can then be defined as $\mathbb{P}_{r,t}(X) = \mathbb{P}(d_r = 1 | X, d_r + C = 1)$, which represents the probability of being first treated in period r, conditional on observable characteristics and on either belonging to cohort r or being never-treated.[5]

We can set up a potential outcomes framework. Specifically, as in Callaway and Sant'Anna (2021), I will combine the dynamic potential outcomes framework (Robins, 1986, 1987) with the dynamic treatment adoption setting (Heckman et al., 2016). For $r = q, \ldots, T$, let $y_{i,t}(r)$ represent the potential outcome for unit i at time t had the policy been introduced by period r (i.e., $d_{i,r} = 1$). Whereas $y_{i,t}(0)$ will denote the analogous in period t had the treatment not been received (that is, had the unit never been treated). Then, one can rewrite the observed outcome for a generic unit as follows:

$$y_{i,t} = y_{i,t}(0) + \sum_{r=q}^{T} d_{i,r} \cdot (y_{i,t}(r) - y_{i,t}(0))$$

[3] If always-treated units exist, these will be dropped from the analysis, as there is no pre-treatment period for these units. For an analysis of the case where the treatment is binary, the parameter of interest is the ATT, and the treatment switches on and off at different points in time, please refer to De Chaisemartin and d'Haultfoeuille (2022).

[4] See Ciaccio (2023) for an empirical application in which never-treated units serve as a more suitable comparison group, as some later treated units modified their behavior to delay the policy's introduction. As discussed in the next section, this would constitute a violation of the Limited Treatment Anticipation assumption, which is necessary for identifying the estimands considered in this paper.

[5] When using not-yet-treated units as the comparison group, the generalized propensity score takes the form $\mathbb{P}_{r,s}(X) = \mathbb{P}(d_r = 1 | X, d_r + (1 - D_s)(1 - d_r) = 1)$. Here, $\mathbb{P}_{r,s}(X)$ denotes the probability of being first treated in period r , conditional on observable characteristics and on either belonging to cohort r (i$d_r = 1$) or being in the not-yet-treated group by time s ($(1 - D_s)(1 - d_r) = 1$).

That is, we will observe only one of the two mutually exclusive potential outcomes for each unit, and the realization of what we will observe will depend on whether unit i is receiving the treatment or not. Specifically, the observed y will be the untreated potential outcome, $y(0)$, for units that never participate in the treatment (true for all $t = 1, \ldots, T$). On the other hand, for units first treated in period r, the observed outcome will be $y(0)$ for all $t = 1, \ldots, r-1$, whereas y will be equal to $y(r)$ for all $t \geq r$. The fact that data will reveal only one of the two mutually exclusive quantities (either $y(r)$ or $y(0)$) constitutes a selection problem. This selection problem represents the *fundamental problem of causal inference* (Holland, 1986).

I will also make the following assumption for the remainder of the paper:

Assumption 2 (*Random Sampling*). $\{Y_{i,1}, Y_{i,2}, ..., Y_{i,T}, X_i, D_{i,1}, D_{i,2}, ..., D_{i,\tau}\}_{i=1}^{n}$ *is independent and identically distributed.*

This assumption requires the availability of panel data. In Appendix B, I extend the analysis to settings where repeated cross-sections, rather than panel data, are available. To do so, I follow Callaway et al. (2018) and Callaway and Sant'Anna (2021).

Assumption 2 is a common assumption in the Diff-in-Diff literature. It allows us to introduce uncertainty in our setting and consider potential outcomes as random variables. Specifically, Assumption 2 implies that any unit i is the realization from a super-population of interest. It is worth stressing that, however, Assumption 2 neither rules out any time series dependence nor imposes any restriction between the relation between $D_{i,t}$ and $(y(0), y(r))$.

Under Assumptions 1 and 2, it is possible to define the estimand of interest. Furthermore, it is worth emphasizing that assuming random sampling rules out the possibility of interference–commonly referred to as the *no interference* assumption. This means that the potential outcome of unit i is not affected by the treatment status of unit $j \neq i$.

The no interference assumption, together with the assumption of a well-defined treatment (i.e., *treatment consistency*) constitute the *Stable Unit Treatment Value Assumption (SUTVA)*.[6]

Before introducing the assumptions needed for identification, I will now introduce the notation both for the quantile and the tests of stochastic dominance ranking. This will be crucial since we are interested in understanding the impact of the policy along with the distribution of our outcome of interest. For $\tau \in [0, 1]$ the τth quantile q_τ of a generic random variable T is defined as:

$$\mathbb{P}(T \leq q_\tau) = F_T(q_\tau) = \tau$$

or analogously as $q_\tau = F_T^{-1}(\tau) := inf\{t : F_T(t) \geq \tau\}$. Where, for instance, $\tau = .25$ represents the 25^{th} percentile.

However, summarizing the distributional treatment effects by comparing quantiles via a quantile-by-quantile approach relies implicitly on the rank invariance assumption. That is, treated and never-treated units with the same rank also have the same unobserved characteristics. As Maasoumi and Wang (2019) show, empirical evidence usually dismisses this assumption.

To overcome this issue, I advocate for the reader to undertake another approach that first summarizes the impact by an appropriate evaluation function and then takes the difference between the

[6]For a discussion on potential violations of SUTVA and approaches to account for spillover effects, see Butts (2024); Fiorini et al. (2024); Xu (2025).

two (evaluation functions), as suggested by Maasoumi and Wang (2019). This approach contrasts with the quantile-by-quantile approach, which requires computing quantile-by-quantile and then taking the difference between the quantile for the treated and untreated potential outcomes.

There is no unique definition of the distribution's evaluation function: "averages, inequality measures, and entropies are all well-known functions of distributions that summarize its quantiles anonymously, without regard to the identity of those who occupy a given quantile. Each function attributes its own weights to different wage levels." (Maasoumi and Wang, 2019, p. 2439). Further, since there are many valid alternatives whose choice often relies on a decision-theoretic framework, I adapt to the framework currently under analysis the statistical tests for stochastic dominance rankings proposed in Maasoumi and Wang (2019).

Considering partial orderings between random variables has two main advantages besides not relying on rank invariance. First, there is no reason to prefer one inequality measure over another: often, choosing one over the other hinges upon a decision-theoretic framework. Tests for stochastic dominance rankings overcome this problem as they tell which of the two distributions dominates, irrespective of a specific weighting scheme.

Second, when the distributions for treated and untreated potential outcomes intertwine, the researcher may reach different conclusions depending on the inequality measure considered. The ranking of these will differ depending on the underlying evaluation functions. This is particularly crucial, even when rank invariance holds. When the two distributions intersect, the treatment effect changes sign, and so the conclusion drawn by the researcher will strongly rely on the measure considered. In this case, stochastic dominance tests also provide a more robust approach to carrying out policy evaluation, as they assess whether distributions can be consistently ranked across a broad range of evaluation functions with a certain level of statistical confidence.

Let us now define what tests of stochastic dominance rankings are. I will use the notation used by Maasoumi and Wang (2019). Let U_1 be the class of von Neumann-Morgenstern utility functions u that are increasing (i.e., $u' \geq 0$) in the potential outcome of interest, $y(d)$ with $d = \{0, 1\}$. Let U_2 be the class of utility function in U_1 such that the utility function is concave ($u'' \leq 0$).

First-order Stochastic Dominance – The treated potential outcome, $y_t(r)|d_r = 1$, first-order stochastically dominates (denoted by $y_t(r)|d_r = 1$ FSD $y_t(0)|d_r = 1$) the untreated potential outcome for the treated, $y_t(0)|d_r = 1$, if and only if

1. $\mathbb{E}u(y_t(r)|d_r = 1) \geq \mathbb{E}u(y_t(0)|d_r = 1)$ for all $u \in U_1$ with the inequality being strict for some u, or
2. $F_{y_t(r)|d_r=1}(y) \leq F_{y_t(0)|d_r=1}(y)$ for all y with the inequality being strict for some values of y, or
3. $y_{t,\tau}(r)|d_r = 1 \geq y_{t,\tau}(0)|d_r = 1$ for all points in the support of y.

where $(y_{t,\tau}(r)|d_r = 1) - (y_{t,\tau}(0)|d_r = 1)$ represents the difference in the treated and the untreated potential outcomes for the treated group at the τth quantile.

Second-order Stochastic Dominance – The treated potential outcome, $y_t(r|d_r = 1)$, second-order stochastically dominates (denoted by $y_t(r)|d_r = 1$ SSD $y_t(0)|d_r = 1$) the untreated potential outcome for the treated, $y_t(0)|d_r = 1$, if and only if

1. $\mathbb{E}u(y_t(r)|d_r = 1) \geq \mathbb{E}u(y_t(0)|d_r = 1)$ for all $u \in U_2$ with the inequality being strict for some u, or

2. $\int_{-\infty}^{y} F_{y_t(r)|d_r=1}(z)dz \leq \int_{-\infty}^{y} F_{y_t(0)|d_r=1}(z)dz$ for all y with strict inequality for some values of y, or

3. $\int_0^{\tau} (y_{t,u}(r)|d_r = 1)\, du \geq \int_0^{\tau} (y_{t,u}(0)|d_r = 1)\, du$ for all points in the support of y.

If, for instance, $y_t(r)|d_r = 1$ FSD $y_t(0)|d_r = 1$, then it is also true that $y_t(r)|d_r = 1$ SSD $y_t(0)|d_r = 1$. This is because first-order stochastic dominance implies second-order stochastic dominance.

2.2.2 The cohort-time distributional treatment effect on the treated

The fact that, as Manski and Pepper (2013) notice, data will reveal only one of the two potential outcomes for the same unit at the same time ($y_{i,t}(r)$ or $y_{i,t}(0)$), makes challenging finding an appropriate way to recover the counterfactual outcome - the outcome one would observe had the policy not been implemented.

Researchers often employ different strategies to retrieve counterfactual outcomes and evaluate the impact of the policy of interest. In non-experimental settings, one of the most common causal effect parameters is the average effect for the treated group, also known as the Average Treatment Effect on the Treated (ATT). Using the potential outcomes framework, the ATT can be written as $ATT = \mathbb{E}(y(1) - y(0)|d = 1)$ (where d is a dummy taking value 1 if the unit is treated).

Since, for treated units, we do not observe the counterfactual outcome, researchers commonly use different assumptions to retrieve a consistent estimator of $y(0)$.[7] One of the most commonly used is the Parallel Trends (PT) assumption (Heckman et al., 1998). In the canonical DiD setting with two periods and two groups, the PT assumption requires that had the policy not been implemented, the average outcome evolution for the treated and untreated units would have remained stable over time (i.e., $\mathbb{E}(y(1) - y(0)|d = 1) = \mathbb{E}(y(1) - y(0)|d = 0)$). Different studies have recently generalized this assumption to cases where multiple periods and multiple time groups are available (Borusyak et al., 2021; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Wooldridge, 2021; De Chaisemartin and d'Haultfoeuille, 2022).

Sometimes, however, policymakers may be concerned with analyzing the impact of a specific policy on specific parts of the distribution of the outcomes considered besides the average effect. When the treatment is likely to be heterogeneous along the distribution of our outcome of interest, considering the ATT alone might be misleading. In such cases, it becomes essential to analyze the distributional impact of the policy for a sound policy evaluation. In such cases, researchers commonly compare distributions of treated and untreated potential outcomes (e.g., see Carneiro et al., 2003).

Similar to what researchers do to evaluate the average causal impact of policy interventions, when the effect is likely to be heterogeneous, the ATT's analogous (distributional) causal parameter of interest is the Quantile Treatment Effect on the Treated. While extensive literature exists on how to retrieve an estimator of the ATT – also in non-standard Difference-in-Differences settings – only a few papers tackle the issue of how to retrieve an estimator of the quantile treatment effects (Athey and Imbens, 2006; Bonhomme and Sauder, 2011; Callaway et al., 2018; Callaway and Li, 2019). Specifically, to the best of my knowledge, no estimators exist to estimate the distributional impact of a policy in a context with staggered treatment adoption.

[7] Please refer to Imbens and Wooldridge (2009) for a review of these assumptions.

However, as said in the previous section, comparing quantiles hinges upon the implicit assumption of rank invariance. For this reason, it is preferable to consider other approaches to summarize the distributional treatment effects that do not require specifying which units occupy a specific quantile.

Therefore, this paper aims to provide a method to recover the whole distribution of untreated potential outcomes for the treated group in a context with multiple groups and periods with variations in treatment timing. To reach this goal, I generalize the estimator proposed in Callaway and Li (2019) by building on the idea behind the group-time average treatment effect proposed by Callaway and Sant'Anna (2021).

Once the entire counterfactual distribution is obtained, the researcher can choose whether to summarize the impact of the policy along the distribution of the outcome of interest by exploiting a quantile-by-quantile approach or first summarizing it by an appropriate evaluation function and then taking the difference between the two evaluation functions, as suggested by Maasoumi and Wang (2019). For completeness, I will provide both an estimator of the quantile treatment effects on the treated and an estimator of the distributional treatment effects on the treated based on tests of stochastic dominance rankings. Borrowing the name from Callaway and Sant'Anna (2021), I call these causal parameters as *cohort-time distributional treatment effects on the treated.*

Specifically, if we denote with $F_{y_t(r)|d_r=1}$ and $F_{y_t(0)|d_r=1}$ the distribution of treated and untreated potential outcomes for units that first receives the treatment in period r, then the QTT parameter for units first treated in group r evaluated at any given time period $t \geq r$ can be defined as follows:

$$QTT_{r,t}(\tau) = F^{-1}_{y_t(r)|d_r=1}(\tau) - F^{-1}_{y_t(0)|d_r=1}(\tau) \quad \tau \in [0,1], r = q, \ldots, T, \text{ and } t \geq r, \tag{2.1}$$

Besides allowing heterogeneity of the treatment effect along with the distribution of y, the parameter $QTT_{r,t}(\tau)$ allows additional heterogeneity across treated cohorts (note that, for example, even keeping t and τ fixed, the QTT experienced by cohort r may be different from that experienced by cohort r') and over time. Moreover, the researcher can employ the *cohort-time quantile treatment effects* to learn about the overall impact of the policy by constructing an overall quantile treatment effect or to highlight heterogeneity along specific dimensions (such as, how does the QTT vary with the length of exposure to the treatment).

As far as the distributional treatment effect based on tests of stochastic dominance rankings is concerned, using the definition of stochastic dominance, one can check for each post-treatment period t and each treated cohort r whether $y_t(r)|d_r = 1$ FSD $y_t(0)|d_r = 1$ or $y_t(r)|d_r = 1$ SSD $y_t(0)|d_r = 1$. As pointed out in the previous subsection, comparing distributions rather than units – besides overcoming the issue of assuming rank invariance – allows the researcher to have a clearer picture of the overall effect of the policy along the distribution of y. This becomes particularly crucial when treated and untreated potential outcomes distributions for the treated group cross, making policy evaluation based on a quantile-by-quantile approach extremely sensitive to the quantile analyzed.

2.2.3 Identifying assumptions

If the Parallel Trends assumption was enough to retrieve a consistent estimator of the ATT in the simple (2×2) DiD scenario – and under a generalization of this assumption to the multi-period setting, numerous studies proved to lead to the identification of ATT even in much more complicated settings – identification of the distributional treatment effects is more challenging. A simple distributional version

of the PT will no longer be sufficient (Fan and Yu, 2012).

Since $F_{y_t(r)|d_r=1}$ is identified from the data, to reach point-identification of the distributional treatment effect in a setting with staggered treatment adoption, we need to identify $F_{y_t(0)|d_r=1}$, the distribution of untreated potential outcome for units treated in cohort r. To reach identification, let us make the following assumptions:

Assumption 3 (*Limited Treatment Anticipation*). *There is a known* $\rho \geq 0$ *such that* $\mathbb{P}\{y_t(r) \leq y|X, d_r = 1\} = \mathbb{P}\{y_t(0) \leq y|X, d_r = 1\}$ *a.s. for all* $r = q, \ldots, T$ *and* $t = 1, \ldots, T$ *such that* $t < r - \rho$.

Assumption 3 is an extension from the mean to the entire distribution of the Limited Treatment Anticipation assumption commonly invoked in DiD settings with staggered treatment adoption (see, for instance, Callaway and Sant'Anna (2021)). It states that, before the treatment occurs, the treated and untreated potential outcomes for "eventually" treated units are the same in the pre-treatment period. In other words, this means that units cannot anticipate the treatment (i.e., $\rho = 0$), or if they can expect it, the time span in which there are possible anticipatory behaviors is known and limited (i.e., $\rho > 0$, and it is such that $t < r - \rho$). Specifically, when $\rho = 0$, it imposes a No Anticipation assumption.

It is worth stressing that Assumption 3 implies that the distributional treatment effect is 0 for all $t < r - \rho$. Further, Assumption 3 is likely to hold in empirical practice when units do not expect the implementation of the policy and/or units cannot "decide" whether to be treated or not.

Next, the following assumption provides a distributional version of the classic Parallel Trends assumption when there is a staggered treatment adoption.

Assumption 4 (*Conditional Distributional PT based on a "Never-treated" Group*). *Let* ρ *be defined as in Assumption 3. For each* $r, t \in \{q, \ldots, T\}$ *such that* $t \geq q - \rho$,

$$\mathbb{P}\left(\Delta y_t(0) \leq \Delta Y|X, d_r = 1\right) = \mathbb{P}\left(\Delta y_t(0) \leq \Delta Y|X, C = 1\right) \quad \text{a.s.}$$

where $\Delta y_t = y_t - y_{t-1}$. Assumption 4 states that, in the absence of treatment, the counterfactual distribution of potential outcomes for units treated in cohort r would have evolved in parallel to that of the never-treated units over time. The Distributional PT Assumption is commonly used in the existing literature to retrieve the counterfactual distribution (Fan and Yu, 2012; Callaway and Li, 2019; Miller, 2023), extending the idea of the common trends assumption from the average to the entire distribution of untreated potential outcomes.

To give the intuition, suppose we want to study the effect of a state-level minimum wage increase on workers' wages. The Distributional PT Assumption implies that, by comparing states that eventually raise the minimum wage with those that never do, the path of earnings, in the absence of the policy, would have been similar between treated and never-treated states. If, for instance, the path of earnings in the absence of the policy depends on education or gender, then the Distributional PT assumption holds only after controlling for these variables. Failing to control for these covariates would lead to biased estimates.

On the other hand, this assumption may fail if different parts of the outcome distribution evolve differently across groups before treatment (e.g., the lower tail of the income distribution in the treated group grows faster than in the control group). For instance, before the minimum wage increase, low-wage workers in treated states may have already experienced faster wage growth due to local labor shortages, while in never-treated states, wages remained relatively stagnant. In this case, even without

the minimum wage increase, the wage distribution in the treated state would have evolved differently from that in the control states. Assuming distributional parallel trends in this scenario could lead to incorrect attribution of pre-existing wage growth or labor market changes to the minimum wage policy, resulting in biased estimates of the treatment effect.[8]

It is important to note that while the common mean PT assumption requires that trends be similar on average between treated and never-treated cohorts, the distributional PT assumption requires that the entire distribution evolves similarly over time.

Unlike selection-on-observables invoked by Firpo (2007),[9] the distributional parallel trends also allow unobservable confounders to vary between treated and untreated cohorts. Allowing for selection on unobservables makes Assumption 4 weaker. It is also weaker than Assumptions 1 and 2 in Bonhomme and Sauder (2011), which require that the unobserved time-varying shock in the pre-treatment period and the time-varying unobserved component in the equation for the untreated potential outcome in the post-treatment period be independent of treatment assignment and independent of unobservable confounders (once controlling for D_i).

As Callaway et al. (2018) point out, if the pre-treatment period is long, then Assumption 4 is testable under the assumption of strict stationarity of the time series of changes in untreated potential outcomes.

Despite being stronger than the mean PT assumption–and more than sufficient to achieve point identification of the ATT–[10] Assumption 4 is no longer sufficient to achieve point identification of $Fy_t(0)|d_r = 1$. Exploiting results on the distribution of the sum of two random variables, Fan and Yu (2012) show that the Distributional PT assumption is insufficient to point-identify the counterfactual distribution for the treated group, making it more challenging to identify the distributional treatment effect. Without additional assumptions, $F_{y(0)|d_r=1}$ can only be partially identified, with potentially wide bounds, as shown by Callaway and Li (2019).

The reason why Assumption 4 is insufficient for point identification of the distributional treatment effect is that different distributions of untreated potential outcomes in period t may appear observationally equivalent. As Callaway et al. (2018) highlight, an additional assumption on the dependence structure (or copula) between $y_{t-1}(0)$ and $\Delta y_t(0)$ is required. Even if the marginal distributions $F_{y_{t-1}(0)|d_r=1}$ and $F_{\Delta y_t(0)|d_r=1}$ can be identified separately, since "observations are observed separately for untreated and treated individuals,..., the joint distribution is not identified" (Callaway and Li, 2019, p.1585).

To retrieve the joint distribution of $y_{t-1}(0)$ and $\Delta y_t(0)$, following Callaway and Li (2019), I will employ Sklar's theorem (Sklar, 1959) which states that any joint distribution can be expressed as a function (or *copula*) of its marginal distributions. Specifically, for any two random variables T and W, Sklar's theorem says that the joint distribution, $F_{T,W}$, can be written as:

$$F_{T,W} = C_{T,W}(F_T(t), F_W(w))$$

where $C_{T,W}(\cdot,\cdot)$ is the copula.[11].

[8] If the assumption fails, bias will vary across quantiles, leading to overestimated effects at some quantiles and underestimated effects at others.

[9] The selection-on-observable assumption–also known in the program evaluation literature as unconfoundness assumption–requires $y(0)$ and $y(1)$ to be independent of treatment assignment, once controlling for a complete set of pre-treatment characteristics.

[10] Once the counterfactual distribution is estimated, the average counterfactual distribution for the treated cohort can be computed as $\mathbb{E}(y(0)|d_r = 1) = \int_0^1 F^{-1}y(0)|d_r = 1(\tau)d\tau$. The ATT can then be straightforwardly retrieved.

[11] The copula captures all the dependence structure between the two random variables but contains no information about

Copulas are widely used in nonparametric statistics because they allow researchers to model dependence between two random variables independently of their marginal distributions. Moreover, "much of the usefulness of copulas in the study of nonparametric statistics derives from the fact that for strictly monotone transformations of the random variables, copulas are either invariant or change in predictable ways" (Nelsen, 2006, p. 25).

The final step is to specify the copula function in this setting. In particular, following Callaway et al. (2018), I will assume that the dependence structure is the same for treated units in cohort r and never-treated units:

Assumption 5 (*Conditional Copula Invariance based on a "Never-treated" Group*). *For all* $x \in \chi$ *and for all* $(u, v) \in [0, 1]^2$

$$C_{\Delta y_t(0), y_{t-1}(0)|X, d_r=1}(u, v) = C_{\Delta y_t(0), y_{t-1}(0)|X, C=1}(u, v)$$

Assumption 5 differs from the assumption in Callaway and Li (2019), who instead impose that the copula between $y_{t-1}(0)$ and $\Delta y_t(0)$ remains stable over time. The key advantage of adopting Conditional Copula Invariance over the Conditional Copula Stability assumption in Callaway and Li (2019) is that it does not require access to panel data with at least two pre-treatment periods. This flexibility allows researchers to work with repeated cross-sections, as I demonstrate in Appendix B.

Specifically, Assumption 5 recovers the missing dependence between $y_{t-1}(0)$ and $\Delta y_t(0)$ by replacing the unknown copula, $C_{\Delta y_t(0), y_{t-1}(0)|X, d_r=1}(\cdot, \cdot)$, with the "observed" dependence from the never-treated group.

Intuitively, in the example given above about the increase in the state-level minimum wage, Assumption 5 implies that the dependence between the change in wages over time and the pre-treatment level of wages is identical between treated and never-treated units in the absence of the policy. If, for example, the dependence between the change in wages over time and the initial level of wages is influenced by the level of education, then for Assumption 5 to hold, we should condition on educational attainment.

Let us now consider a case where Assumption 5 is likely to be violated. Consider again the increase in the state-level minimum wage, and suppose we are interested in evaluating its impact on wages for low-wage workers. Suppose that, by time t, all states have maintained a stable minimum wage for over a decade. For simplicity (though this is not necessary), assume that all states face the same wage level for low-wage individuals, w. Now, suppose that minimum wage is first increased by state s in period $t + 1$. However, in the year prior to this increase, between periods $t - 1$ and t, wages for low-wage workers rise much more in the states that, during the observed period, do not raise the minimum wage (i.e., never-treated states) due to local labor shortages. In contrast, in the eventually later-treated states, wages for low-wage workers remain almost stagnant.

In this context, the copula represents the dependence between the initial wage level w and the subsequent wage increase occurring between periods $t-1$ and t. Since this dependence differs between treated and never-treated states even before the minimum wage increase, Assumption 5 is likely to be violated. Using this method will, therefore, lead to biased estimates. The extent of the bias will depend on the difference in the dependence described above between never-treated and eventually treated units. The larger the difference in dependence, the greater the bias.

their marginal distributions. For a detailed discussion on copulas, see, for example, Nelsen (2006).

It is important to highlight that Assumption 5 imposes no restrictions on the marginal distributions; it only restricts the dependence (i.e., the initial distributions can indeed differ between treated and never-treated groups).

Moreover, as noted by Callaway et al. (2018) and Callaway and Li (2019), Assumption 4 does not imply Assumption 5, nor the reverse is true. The reason the two assumptions are not mutually implied is that while Assumption 4 restricts only the marginals of the change in untreated potential outcomes, Assumption 5 restricts the dependence between $y_{t-1}(0)$ and $\Delta y_t(0)$. To illustrate this, suppose that $y_{t-1}(0)$ and $\Delta y_t(0)$ are jointly normally distributed with means μ_1 and μ_2 respectively, and variances σ_1^2 and σ_2^2, respectively. If the correlation between $y_{t-1}(0)$ and $\Delta y_t(0)$ is denoted as ρ_t, then under Assumption 5 we require that ρ_t be independent of treatment assignment, conditionally on X and either being treated in cohort r or never-treated. This assumption does not restrict the means or variances of the two random variables, but only their correlation.

While no formal tests have been proposed thus far to check whether Assumption 5 hold, Callaway and Li (2019) suggests using rank correlation measures, such as Kendall's τ, to assess its plausibility in the pre-treatment periods. Using a test based on Kendall's τ, they show the validity of their Copula Stability assumption when evaluating the effect of increasing minimum wages on the unemployment rate in the U.S. during years 2000-2007, thus providing evidence of contexts where assumptions like Assumption 5 are likely to hold.

On the one hand, the researcher can inspect whether the dependence between the initial level of outcomes and the change in outcomes over time is similar between treated and never-treated units. However, if such dependence is more likely to be similar among specific groups within the population, the researcher should condition on these characteristics. Conditioning on observable pre-treatment characteristics should increase the likelihood that Assumption 5 holds in practice.

Under Assumption 4 and Assumption 5 we can point-identify $F_{y(0)|d_r=1}$. However, since the copula representation is not unique unless the random variables are continuous (Joe, 1997; Nelsen, 2006), following Callaway et al. (2018) and Callaway and Li (2019), I will impose the following additional assumption:

Assumption 6 (*Continuity*). *The random variables $Y_{t-1}(0)$ and $\Delta Y_t(0)$ have continuous distribution conditional on either being part of the treated cohort r (i.e., $d_r = 1$) or being never-treated (i.e., $C = 1$), and $Y_t(r)|d_r = 1$ is also continuously distributed on its support. Moreover, each of these distributions has a compact support with marginal distributions, which are (uniformly) bounded away from 0 and 1 over their respective support.*

Lastly, the following assumption will be required only to recover $F_{y(0)|d_r=1}$ when conditioning on observable pre-treatment covariates:

Assumption 7 (*Overlap*). *For each $r, t \in \{q, ..., T\}$, there exists some $\varepsilon > 0$ s.t. $\mathbb{P}(d_r = 1) > \varepsilon$ and $\mathbb{P}_{r,t}(X) < 1 - \varepsilon$ a.s.*

where $\mathbb{P}(d_r = 1)$ denotes the probability of being treated in cohort r and $\mathbb{P}_{r,t}(X) = \mathbb{P}(d_r = 1|X, d_r + C = 1)$ denotes the generalized propensity score. Assumption 7 is the common overlap condition and generalizes the assumptions done in Firpo (2007); Callaway et al. (2018); Callaway and Li (2019) to the multiple periods and multiple groups setting. In words, the first part of Assumption 7 states that there is a positive probability of being treated in cohort r (compared to being always never-treated). The second part, instead, says that, for all the combinations of (r, t), the propensity score is bounded away from 1.

2.2.4 Non-parametric identification of the distribution of untreated potential outcome

In this subsection, I will show that, under the aforementioned assumptions, the full counterfactual distribution for the treated group, $F_{y(0)|d_r=1}$, can be non-parametrically point-identified in Difference-in-Differences models with multiple periods and multiple groups.

Suppose absurdly that pre-treatment covariates play no role in the identification. The following theorem shows that under Assumptions 1, 2, 6 and an unconditional version of Assumptions 3–5, we can point identify $F_{y(0)|d_r=1}$.

Theorem 1. *Suppose Assumptions 1, 2, 6, and unconditional version of Assumptions 3-5 hold. Then* $F_{y_t(0)|d_r=1}(\tau)$ *is identified:*

$$\begin{aligned} F_{y_t(0)|d_r=1} &= \mathbb{P}\left(y_t(0) \leq y | d_r = 1\right) \\ &= \mathbb{E}\Big[\mathbb{1}\left(\Delta_{[r-\rho-1,t]} y(0)\right) \leq y - \\ &\quad F^{-1}_{y_{r-\rho-1}(0)|d_r=1}\left(F_{y_{r-\rho-1}(0)|C=1}\left(y_{r-\rho-1}(0)\right)\right)\Big] \,|C=1\Big] \end{aligned}$$

where $\Delta_{[r-\rho-1,t]} y(0) = y_t(0) - y_{r-\rho-1}(0)$ represents the long difference.

Theorem 1 is the main result of this paper. It shows that, under the classical assumptions done in the staggered Diff-in-Diff literature, an extension of the common PT assumption to the entire distribution of change in untreated potential outcomes and a "new" assumption regarding the joint distribution of $y_{t-1}(0)$ and $\Delta y_t(0)$, one can point reach point-identification of $F_{y_t(0)|d_r=1}(\tau)$. Once $F_{y_t(0)|d_r=1}(\tau)$ is identified, we can also identify $F^{-1}_{y_t(0)|d_r=1}(\tau)$.

Theorem 1 implies that units belonging to treated cohort r must be similarly distributed to never-treated units in terms of both marginals distributions of $y(0)_{t-1}$ and $\Delta y(0)_t$, but also in terms of the dependence between these marginals. This is guaranteed by Assumptions 4 and 5.[12]

The intuition lying behind the results in Theorem 1 is the following – and is almost identical to that given by Callaway and Li (2019) for their main identification results. Note that

$$\mathbb{P}\left(y_t(0) \leq y | d_r = 1\right) = \mathbb{E}\left[\mathbb{1}\left(\Delta_{[r-\rho-1,t]} y(0) + y_{r-\rho-1}(0)\right) \leq y | d_r = 1\right]$$

is an integral (remember that y is continuous) over the joint distribution between the pre-treatment level in the untreated potential outcome, $y_{r-\rho-1}$, and the change in untreated potential outcome. This joint distribution can be identified under Assumption 5, which allows recovering the missing dependence by replacing the unknown copula for the treated with that observed for the never-treated. However, since $\Delta_{[r-\rho-1,t]} y(0)$ is not observed for the treated group r, we need Assumption 4 to replace $F^{-1}_{\Delta_{[r-\rho-1,t]} y(0)|d_r=1}(\cdot)$ with $F^{-1}_{\Delta_{[r-\rho-1,t]} y(0)|C=1}(\cdot)$. For the entire proof, please refer to Appendix A.

The following example shows the conditions for a classic two-way fixed effects regression to satisfy the assumptions needed to identify the group-time average treatment effects proposed by Callaway and Sant'Anna (2021) and the distribution of the untreated potential outcome presented in this paper.

`Example 1.` Consider the following two-way fixed effects regression for the untreated potential

[12] It is worth stressing that, under Assumptions 1-4, it is also possible to identify the ATT in a context with multiple periods and groups with staggered treatment adoption.

outcome:

$$y_{it}(0) = \alpha_t + \eta_i + u_{it}$$

where α_t represents the time-fixed effects, η_i is the unobserved heterogeneity – which may be distributed differently between treated cohorts and never-treated units –, and u_{it} represents the time-varying unobservable shock. Suppose we have access to panel data and assume, for simplicity, there is no treatment anticipation (i.e., $\rho = 0$). For this DGP to satisfy the assumptions of the estimator of the ATT presented in Callaway and Sant'Anna (2021), a sufficient condition for the mean version of Assumption 4 to hold would be that $\mathbb{E}(\Delta u_{it}|d_r = 1) = \mathbb{E}(\Delta u_{it}|C = 1)$. Instead, for the above DGP to satisfy the assumptions of the method presented in this paper, then two sufficient conditions are i) $\Delta u_{it} \perp\!\!\!\perp D$, ii) $C_{\Delta u_t, u_{t-1}|d_r=1} = C_{\Delta u_t, u_{t-1}|C=1}$. Note that conditions i) and ii) allow for the distribution of the time-varying shock to vary over time – thus allowing for serial correlation – as well as for the distribution of u_{it} to be potentially correlated with η_i, in contrast to Bonhomme and Sauder (2011).

The validity of the claims in Example 1 follows the same logic as the proof provided in Appendix A for Example 1 in Callaway and Li (2019). The only difference is that I rely on a copula invariance assumption, while Callaway and Li (2019) rely on a copula stability assumption. Furthermore, note that for the method presented in this paper to work, it is sufficient only to impose restrictions on how the distribution of untreated potential outcomes is generated. Nothing is said about how the treated potential outcome is generated.

However, it is highly unlikely that any of Assumptions 3-5 hold unconditionally. There is a wide literature that shows that, in most applications, these assumptions hold only after having conditioned on pre-treatment observable characteristics (for a discussion, see Heckman et al., 1998; Abadie, 2005).

In the remainder of this section I, therefore, consider the case where I require Assumptions 3–4 to hold conditioning on covariates. Specifically, I will distinguish among two scenarios:

(i) Assumptions 3 and 4 hold after conditioning on covariates, whereas Assumption 5 hold unconditionally;

(ii) Assumptions 3–5 hold after conditioning on covariates.

Let us consider first scenario (i). Then, the only part of Theorem 1 we need to modify to reach identification of the counterfactual distribution is the identification of $F_{\Delta y(0)|d_r=1}$, once having accounted for pre-treatment observable characteristics. Rather than obtaining a conditional version of this distribution function, following Callaway and Li (2019), I will generalize the propensity score method proposed by Firpo (2007) to the staggered treatment adoption scenario.

Despite requiring a parametric specification for $\mathbb{P}_{r,t}(X)$ comes at the additional cost of imposing a distributional form for which there is no guarantee to be correct, this is a commonly used technique that is easy to implement (Abadie, 2005): in the first step one obtains an estimator of the propensity score; in a second stage, one plugs the estimated propensity score in the sample counterparts of $F_{y(0)|d_r=1}$.[13] It is worth stressing that the remaining part of Theorem 1 will continue to be valid thanks to Assumption 5 holding unconditionally.

[13] A way to overcome the fact that this method strongly relies on the correct specification of the functional form of $\mathbb{P}_{r,t}(X)$ would be to use a doubly robust estimator for the $F_{y(0)|d_r=1}$ as the one proposed by Miller (2023). However, this goes beyond the scope of this paper, whose aim is to propose a method to estimate $F_{y(0)|d_r=1}$ in a setting characterized by a staggered intervention.

Proposition 1. *Under Assumptions 1-4, 6, 7, and unconditional version of Assumptions 5 hold,* $F_{y_t(0)|d_r=1}(\tau)$ *is identified:*

$$\begin{aligned} F_{y_t(0)|d_r=1} =& \mathbb{E}\Big[\mathbb{1}\Big[F^{p,-1}_{(y_t(0)-y_{r-\rho-1}(0)|d_r=1)}\Big(F_{(y_t(0)-y_{r-\rho-1}(0)|C=1)}\left(y_t\left(0\right)-y_{r-\rho-1}\left(0\right)\right)\Big) \\ & \leq y - F^{-1}_{y_{r-\rho-1}(0)|d_r=1}\Big(F_{y_{r-\rho-1}(0)|C=1}\left(y_{r-\rho-1}(0)\right)\Big)\Big]\,|C=1\Big] \end{aligned}$$

where

$$F^{p}_{\Delta r-\rho-1,t(0)|d_r=1}(\delta) = \mathbb{E}\left[\frac{C}{p_r}\frac{p_r(x)}{1-p_r(x)}\mathbb{1}\{y_t - y_{r-\rho-1}\leq\delta\}\right] \tag{2.2}$$

which is identified.

Again, the result just obtained is almost identical to the one derived in Theorem 1. The only part that changes is that now $F_{\Delta y(0)|d_r=1}$ is identified by the reweighted distribution (in Eq. (2.2)) of the change in the untreated potential outcome that occurs for never-treated units. This reweighting procedure is almost identical to the one proposed by Abadie (2005) and Firpo (2007) with some minor changes. For a sketch of the proof of why the result obtained in Theorem 1 is still valid, please refer to Appendix A.

The following example shows the importance of Assumption 3 and Assumption 4 holding conditionally in the empirical application.

`Example 2.` Consider the following two-way fixed effects regression for the untreated potential outcome:

$$y_{it}(0) = \alpha_t + \eta_i + x'_{it}\beta + u_{it}$$

Where X is a full set of covariates, which can be distributed differently between treated and never-treated units; $u_{it} = \rho u_{i,t-1} + \varepsilon_{it}$, with $\varepsilon \sim WNN$ process and $|\rho| < 1$. For this DGP to satisfy the assumptions of the model presented in this paper, then two sufficient conditions are i) $\Delta u_{it} \perp\!\!\!\perp D|X$, ii) $C_{\Delta u_t, u_{t-1}|d_r=1} = C_{\Delta u_t, u_{t-1}|C=1}$.

The last scenario I consider in this subsection is the case where also Assumption 5 holds after conditioning on covariates (scenario (ii)). Specifically, the following proposition shows that it is still possible to identify the $F_{y_t(0)|d_r=1}(\tau)$:

Proposition 2. *Suppose that the random variables* $y_{t-1}(0)$ *and* $\Delta y_t(0)$ *are continuosly distributed conditionally on* x*, and that this is true for all* $x \in \chi$ *and both group* r *or the never-treated units. Then, under Assumptions 1-7,* $F_{y_t(0)|d_r=1}(\tau)$ *is identified:*

$$\begin{aligned} \mathbb{P}\left(y_t\left(0\right)\leq y|X=x,d_r=1\right) =& \mathbb{E}\Big[\mathbb{1}\Big[F^{-1}_{\Delta_{r-\rho-1,t}y(0)|X,C=1}\Big(F_{\Delta_{r-\rho-1,t}y(0)|X,C=1}\left(\Delta y_{t,r-\rho-1}\left(0\right)\right)|X\Big)\leq y- \\ & F^{-1}_{y_{r-\rho-1}(0)|X,d_r=1}\Big(F_{y_{r-\rho-1}(0)|X,C=1}\left(y_{r-\rho-1}(0)\right)|X,\Big)\Big]\,|X=x,C=1\Big] \end{aligned}$$

and

$$\mathbb{P}\left(y_t\left(0\right)\leq y|d_r=1\right) = \int_{\chi}\mathbb{P}\left(y_t\left(0\right)\leq y|X=x,d_r=1\right)dF_{X|d_r=1}(x)$$

The result obtained in Proposition 2 is very close to that obtained above in Theorem 1. The only difference compared to Theorem 1 is that one needs to compute first conditional distributions of untreated potential outcomes. Next, the researcher has to integrate out covariates from $\mathbb{P}\left(y_t\left(0\right)\leq y|X=x,d_r=1\right)$ to obtain $\mathbb{P}\left(y_t\left(0\right)\leq y|d_r=1\right)$, the unconditional distribution of untreated potential outcome.

The cohort-time quantile treatment effect

We can construct different causal parameters once we have identified $F_{y(0)|d_r=1}$. Suppose rank invariance holds, then it is possible to show that, under the same assumptions required for the results obtained in Theorem 1 or in Proposition 1 to be valid, one can obtain the inverse of $F_{y(0)|d_r=1}$ ($F^{-1}_{y(0)|d_r=1}(\tau)$ for each $\tau \in (0,1)$). At this point, the *cohort-time quantile treatment effects* are also identified:

$$QTT_{r,t,\rho}(\tau) = F^{-1}_{y_t(r)|d_r=1}(\tau) - F^{-1}_{y_t(\infty)|d_r=1}(\tau) \quad \forall r,t \in \{q,\dots,T\}, t \geq r-\rho.$$

If, instead, one relies on the result obtained in Proposition 2. After $\mathbb{P}\left(y_t\left(0\right) \leq y | X = x, d_r = 1\right)$ is identified, one can invert it, and the conditional cohort-time quantile treatment effects are identified:

$$QTT_{r,t,\rho}(\tau;x) = F^{-1}_{y_t(r)|X,d_r=1}(\tau|x) - F^{-1}_{y_t(0)|X,d_r=1}(\tau|x) \quad \forall r,t \in \{q,\dots,T\} \text{ and } t \geq r-\rho$$

Or one can integrate out covariate from $\mathbb{P}\left(y_t\left(0\right) \leq y | X = x, d_r = 1\right)$, obtain $\mathbb{P}\left(y_t\left(0\right) \leq y | d_r = 1\right)$, invert it and then get:

$$QTT_{r,t,\rho}(\tau) = F^{-1}_{y_t(r)|d_r=1}(\tau) - F^{-1}_{y_t(0)|d_r=1}(\tau) \quad \forall r,t \in \{q,\dots,T\} \text{ and } t \geq r-\rho$$

Which is identified.

Once the cohort-time quantile treatment effects are identified, one can aggregate these parameters, for instance, to highlight heterogeneity along specific dimensions (such as how the treatment effects vary with the length of exposure to the treatment). Or, analogously, one can aggregate these causal parameters to summarize the overall treatment effect in one unique parameter. Specifically, I advocate for the researcher to use aggregations schemes of the cohort-time quantile treatment effects similar to those suggested by Callaway and Sant'Anna (2021).

Aggregation schemes are particularly crucial when there are many parameters to estimate, and the researcher wants to have a quick tool to summarize results to highlight heterogeneity over time or by treatment cohorts. Note, however, that while in the scenario considered in Callaway and Sant'Anna (2021), it made sense to propose a simple weighted average of their treatment effect parameters to summarize the effect of the policy; in principle, in the scenario considered in this paper, one might think of more complicated aggregation schemes to highlight heterogeneity also along the distribution of y. For instance, the researcher might be interested in the policy's overall cumulative effect on the lower tail of the distribution (e.g., evaluate the impact of the introduction of minimum wages on those units that reside in the first two deciles of the income distribution). Nonetheless, to keep things simple and to avoid including an additional dimension of heterogeneity, I will limit myself to generalizing the discussion in Callaway and Sant'Anna (2021) to the QTT case.

Specifically, in the previous section, we focused on stating under which conditions one can identify the QTT parameters. It may be relevant for a policymaker to understand, for instance, the long-run effects of the policy or to understand whether there exist different patterns according to when a unit is first treated. Following Callaway and Sant'Anna (2021), I will propose aggregation schemes of the following form:

$$\theta(\tau) = \sum_{t=q}^{T} \sum_{r=q}^{T} w(r,t) \cdot QTT_{r,t,\rho}(\tau) \quad r \in \{q,\dots,T\} \tag{2.3}$$

where $w(r,t)$ are the researcher's weighting functions. One can address multiple policy-relevant questions by changing $w(r,t)$, such as: *'How does the QTT vary across groups?'*; *'How does the QTT vary with the length of exposure to the treatment?'*; *'What is the cumulative distributional treatment effect of the policy across all groups until time t?'*; or *'What is the overall impact of the policy'*. For instance, if the researcher is interested in understanding how the QTT varies with the length of exposure to the treatment, by denoting with $e = t - r$ the event time (i.e., the time elapsed since the unit first received the treatment), one possible aggregation scheme to highlight the heterogeneity of the QTT with respect to the event time is:

$$\theta^{e}_{exp}(\tau) = \sum_{r=q}^{T} \mathbb{1}\{r + e \leq T\} \mathbb{P}\left(d_r = 1 | r + e \leq T\right) QTT_{r,r+e,\rho}(\tau) \quad r \in \{q, \dots, T\}$$

$\theta^{e}_{exp}(\tau)$ represents the effect – for units belonging to the τ^{th} quantile of the distribution of y– of participating in the treatment for e periods after treatment implementation. $\theta^{e}_{exp}(\tau)$ is computed across all units, which are exactly e periods away from when the policy was first implemented. $\theta^{e}_{exp}(\tau)$ is a common parameter of interest in event studies where the researcher is interested in understanding the dynamic impact of the policy (e.g., see Sun and Abraham, 2021; De Chaisemartin and d'Haultfoeuille, 2022).

If, on the other side, the researcher is interested in evaluating the overall impact of the policy on those units that belong to the τ^{th} quantile, then a straightforward way to obtain an overall effect parameter is the following:

$$\theta^{o}_{weight}(\tau) = \frac{1}{\kappa} \sum_{t=q}^{T} \sum_{r=q}^{T} \mathbb{1}\{t \geq r\} \mathbb{P}\left(d_r = 1 | r \leq T\right) QTT_{r,t,\rho}(\tau) \quad r \in \{q, \dots, T\}$$

where $\kappa = \sum_{t=q}^{T} \sum_{r=q}^{T} \mathbb{1}\{t \geq r\} \mathbb{P}\left(d_r = 1 | r \leq T\right)$. $\theta^{o}_{weight}(\tau)$ is a weighted average of QTTs putting more weight on those QTTs whose size is larger. However, one drawback of $\theta^{o}_{weight}(\tau)$, as also Callaway and Sant'Anna (2021) note, is that it systematically weights those groups that participate for a longer period in the treatment.

Tests of stochastic dominance rankings

If rank invariance is unlikely to hold in the context analyzed, one can perform statistical tests of stochastic dominance rankings or use inequality measures. The same reasoning applies when the two distributions cross – the $QTT_{r,t,\rho}(\tau)$ changes sign when going from $\tau = 0$ to $\tau = 1$ – as policy implications based on these results are sensitive to the choice of the quantiles considered.

For the remainder of this paper, I will consider only tests of stochastic dominance rankings. Once the counterfactual distribution is identified, one can construct tests of stochastic dominance rankings for the treated group r by comparing the distribution of the treated and untreated potential outcomes for the treated (i.e., $F_{y_t(r)|d_r=1}$ and $F_{y_t(0)|d_r=1}$). Specifically, once the entire counterfactual distribution is identified, the researcher can invert it and obtain all the possible values of $y_t(0)|d_r = 1$ to compute tests of this type.

For instance, when considering first-order stochastic dominance, we will say that the treated potential outcome, $y_t(r)|d_r = 1$, first-order stochastically dominates the untreated potential outcome for the treated, $y_t(0)|d_r = 1$, if and only if

1. $\mathbb{E}u(y_t(r)|d_r = 1) \geq \mathbb{E}u(y_t(0)|d_r = 1)$ for all $u \in U_1$ with the inequality being strict for some u or

2. $F_{y_t(r)|d_r=1}(y) \leq F_{y_t(0)|d_r=1}(y)$ for all y with strict inequality for some values of y, or

3. $y_t(r)|d_r = 1 \geq y_t(0)|d_r = 1$ for all points in the support of y.

Again, this is possible as we have identified the entire counterfactual distribution.

2.3 Estimation and Inference

In the previous sections, we have focused on identifying the counterfactual distribution of the untreated potential outcome for the treated group. Once identified, we showed that one can build different causal parameters. Moreover, when considering the QTT, the aggregations schemes suggested by Callaway and Sant'Anna (2021) can be easily generalized. In this section, I will propose a way to estimate these parameters and different ways to conduct valid inference according to the type of estimand considered.

Suppose one is interested in estimating the parameters presented in (2.1). Using the analogy principle (Manski, 1994), one straightforward way to construct an estimator of these parameters is to use empirical distribution functions (EDFs), as Callaway and Li (2019) do.[14]

In the case in which covariates play no role (the case analyzed in Theorem 1), we can estimate the cohort-time quantile treatment effects by:

$$\widehat{QTT}_{r,t,\rho}(\tau) = \hat{F}^{-1}_{y_t(r)|d_r=1}(\tau) - \hat{F}^{-1}_{y_t(0)|d_r=1}(\tau)$$

where $\widehat{F}^{-1}_{y_t(r)|d_r=1}(\tau)$ is directly estimated from the data. Specifically, the researcher first estimates $\widehat{F}_{y_t(r)|d_r=1}(y)$, and then obtains the quantile function for the required quantile of interest by simply inverting $\widehat{F}_{y_t(r)|d_r=1}(y)$. That is:

$$\hat{F}^{-1}_{y_t(r)|d_r=1}(\tau) = inf\{y : \hat{F}_{y_t(r)|d_r=1}(y) \geq \tau\}$$

Whereas we can estimate the counterfactual quantile exploiting the result obtained in Theorem 1, where EDFs and quantile functions will be estimated by their sample analogs:

$$\hat{F}^{-1}_{y_t(0)|d_r=1}(\tau) = inf\{y : \hat{F}_{y_t(0)|d_r=1}(y) \geq \tau\}$$

where

$$\begin{aligned}\hat{F}_{y_t(0)|d_r=1}(y) =& \frac{1}{n_0}\sum_{i\in 0} \mathbb{1}\left[\hat{F}^{-1}_{\Delta_{[r-\rho-1,t]}y|C=1}\left(\hat{F}_{\Delta_{[r-\rho-1,t]}y|C=1}\left(\Delta_{[r-\rho-1,t]}y\right)\right)\right.\\ &\left.\leq y - \hat{F}^{-1}_{y_{r-\rho-1}|d_r=1}\left(\hat{F}_{y_{r-\rho-1}|C=1}\left(y_{r-\rho-1}\right)\right)\right]\end{aligned}$$

where n_0 represents the number of never-treated units in periods $r-\rho-1$ and t.

[14] Under the assumptions required in Theorem 1 and the additional assumption that the support of $y(j)$ is compact (with $j \in \{r, 0\}$), Callaway and Li (2019) show that this estimator is consistent. Moreover, using a functional central limit theorem, they also show that their estimator converges uniformly to a Gaussian process. Their results can be easily extended to this context, as what changes is only how we construct the benchmark but not the estimator itself.

Regarding statistical inference, depending on the context under analysis, I advocate for the researcher to use the most appropriate procedure to carry out valid inference. For instance, when there is no serial correlation, or it is unlikely to have clustered dependence of any sort, one can use the empirical bootstrap proposed in Callaway and Li (2019). The authors prove this procedure leads to uniform confidence bands that cover the $QTT(\tau)$ with fixed probability for all $\tau \in [\varepsilon, 1-\varepsilon] \subset (0,1)$ for some small, positive ε.

Specifically, let $\widehat{QTT}^*(\tau)$ denote an estimate of the QTT using a bootstrapped sample (the steps are identical to those presented above; the only thing that changes is the sample used to form these $\widehat{QTT}^*(\tau)$) and let B be the number of bootstrap iterations and for $b = 1, \ldots, B$. For each iteration, then one needs to compute:

$$I^b = sup_{\tau \in \mathrm{T}} \hat{\Sigma}(\tau)^{-1/2} \mid \sqrt{n} \left(\widehat{QTT}^b(\tau) - \widehat{QTT}(\tau) \right) \mid$$

Where $\hat{\Sigma}(\tau)^{1/2} = (q_{0.75}(\tau) - q_{0.25}(\tau)) \,/\, (z_{0.75}(\tau) - z_{0.25}(\tau))$ represents the interquartile range obtained via bootstrap, divided by the interquartile range of a standard normal random variable. Then one can obtain the $(1-\alpha)$ confidence interval as follows:

$$\hat{C}_{QTT(\tau)} = \widehat{QTT}^*(\tau) \pm c^B_{1-\alpha} \hat{\Sigma}(\tau)^{1/2} / \sqrt{n}$$

Where $c^B_{1-\alpha}$ is the $(1-\alpha)$ quantile of $\{I^b\}_{b=1}^B$.[15] If clustered dependence is likely in the context under analysis, the researcher can adapt the wild cluster bootstrap (or subcluster wild bootstrap (MacKinnon and Webb, 2018) if the number of treated clusters is small) to the QTT case, with minor adaptations. See Cameron and Miller (2015) and MacKinnon et al. (2023) for a review.

As far as the case considered in Proposition 1, as shown in Appendix A, the only additional term that needs to be estimated is $F^{p,-1}_{\Delta_{[r-\rho-1,t]}y(0)|d_r=1}(\delta)$. One can compute it by estimating the following generalization of the estimator proposed by Firpo (2007) and Callaway and Li (2019):

$$\mathbb{F}^p_{\Delta y_{t,r-\rho-1}(0)|d_r=1}(\delta) = \frac{1}{n_{r,t}} \sum_{i \in r} \frac{C}{p_r} \frac{\hat{p}_r(x_i)}{1-\hat{p}_r(x_i)} \mathbb{1}\{\Delta y_{t,r-\rho-1} \leq \delta\} \Bigg/ \frac{1}{n_{r,t}} \sum_{i \in r} \frac{C}{p_r} \frac{\hat{p}_r(x_i)}{1-\hat{p}_r(x_i)}$$

Where $n_{r,t}$ is the number of units used to compute $\widehat{QTT}_{r,t,\rho}(\tau)$, $\hat{p}(X)$ represents an estimator of the propensity score, and the last term in the denominator normalizes the weights to sum to 1 in finite samples (ensuring that $\mathbb{F}^p(\cdot)$ is indeed a distribution function). It is worth stressing that, as both Firpo (2007) and Callaway and Li (2019) show, one can estimate the propensity score parametrically and non-parametrically.

As far as the case considered in Proposition 2 is concerned, it can be computationally demanding. Even if estimated non-parametrically, this would imply estimating five conditional distributions. This may be unfeasible in many empirical applications, especially with $dim\{\boldsymbol{X}\}$ large and n small.

To overcome this issue, one can either estimate the above quantities using quantile regressions or, if covariates are discrete, use the method proposed by Callaway et al. (2018). For instance, Melly and Santangelo (2015) exploit quantile regression to estimate the Change in Changes model proposed by Athey and Imbens (2006) in the context with covariates. On the other side, following Callaway et al.

[15] For extreme quantile, one can use alternative inference procedures (see, Chernozhukov et al., 2016).

(2018), one can estimate, for each possible value of x, $\hat{F}_{y_t(r)|X,d_r=1}$ via its empirical distribution function and then invert it. Analogously, $\hat{F}_{y_t(0)|X,d_r=1}$ can be estimated for each possible values of x exploiting the result in Proposition 2. That is

$$\begin{aligned}\hat{F}_{y_t(0)|X=x,d_r=1}(y) =& \frac{1}{n_{0,x}} \sum_{i \in 0} \mathbb{1}\left[\left(\Delta_{[r-\rho-1,t]} y\right)\right. \\ & \left. \leq y - \hat{F}^{-1}_{y_{r-\rho-1}|X=x,d_r=1}\left(\hat{F}_{y_{r-\rho-1}|X=x,C=1}\left(y_{r-\rho-1}\right)\right) | X = x\right]\end{aligned}$$

Again, one can use empirical distribution functions to estimate each of the quantities in the formula above. Once obtained an estimate of $\hat{F}_{y_t(r)|X,d_r=1}(\tau|x)$ and $\hat{F}_{y_t(0)|X,d_r=1}(\tau|x)$, one can obtain an estimator of the conditional QTT as:

$$QTT_{r,t,\rho}(\tau; x) = F^{-1}_{y_t(r)|X,d_r=1}(\tau|x) - F^{-1}_{y_t(0)|X,d_r=1}(\tau|x) \quad \forall r, t \in \{q, \ldots, T\} \text{ and } t \geq r - \rho$$

for each $\tau \in (0, 1)$ and for each $x \in \chi$.

Once an estimator for the QTT's is obtained, the researcher can aggregate these parameters using the aggregation schemes proposed in the previous section. Using their sample analogs, she can estimate the weighting functions using the analogy principle (Manski, 1994).

If, on the other hand, the researcher is interested in performing tests of stochastic dominance (SD) rankings, one can perform stochastic dominance tests based on the generalized Kolmogorov-Smirnov statistics proposed by Linton et al. (2005). Once we obtain an estimator for $\hat{F}_{y_t(r)|d_r=1}$ and $\hat{F}_{y_t(0)|d_r=1}$ – exploiting the results obtained in Theorem 1, Propositions 1 or 2 (according to the setting under analysis) – tests for FSD or SSD can be obtained as follows:

$$d_{r,t,\rho} = \sqrt{\frac{n_{r,t} \cdot n_{0,t}}{n_{r,t} + n_{0,t}}} \min \sup \left(\hat{F}_{y_t(r)|d_r=1}(y) - \hat{F}_{y_t(0)|d_r=1}(y)\right)$$

$$s_{r,t,\rho} = \sqrt{\frac{n_{r,t} \cdot n_{0,t}}{n_{r,t} + n_{0,t}}} \min \sup \int_{-\infty}^{y} \left(\hat{F}_{y_t(r)|d_r=1}(z) - \hat{F}_{y_t(0)|d_r=1(z)}\right) dz$$

where $n_{r,t}$ and $n_{0,t}$ represent the sample sizes used to estimate $\hat{F}_{y_t(r)|d_r=1}$ and $\hat{F}_{y_t(0)|d_r=1}$, respectively.

Following Maasoumi and Wang (2019), one can construct a test for iid samples based on the pair bootstrap. This procedure allows the researcher to obtain the probability of any of the two SD tests to fall in a specific interval, as well as the $p-$value. For instance, if $\mathbb{P}(d \leq 0)$ is large (e.g., above .90) and d is non-positive, then we can claim that $y_t(r)|d_r = 1$ FSD $y_t(0)|d_r = 1$ with a high degree of statistical confidence (Maasoumi and Wang, 2019).

More sophisticated tests can be built depending on the context under analysis (for a survey on the possible tests that can be used with this approach, see Maasoumi, 2003). Again, depending on the context under analysis, I advocate using the most appropriate test to conduct statistical inference and decide whether the distribution of treated potential outcome at time t for the group first treated in period r stochastically dominates the counterfactual distribution. However, this is beyond the scope of this paper.

2.4 Monte-Carlo Simulations

2.4.1 Monte-Carlo Simulations – varying N and T

In this section, I will analyze the finite-sample properties of the estimators of the distribution of the untreated potential outcome presented in the previous section. Because the results obtained in Section 2.2 do not rely on the specific causal parameter considered, for brevity, I will limit myself to analyzing the estimators' performance in the case in which the researcher is interested in recovering an estimator of the cohort-time quantile treatment effect. Specifically, since the identification results do not depend on specifying the outcome for the treated group, I will assess the finite-sample properties of the methods presented in generating $F_{y_t(0)|d_r=1}$.[16]

In contrast to most existing literature, I will assess the performance of the proposed method in three scenarios: (i) when the data-generating process (DGP) for the untreated potential outcome follows a TWFE regression without covariates (as in Callaway et al., 2018; Callaway and Li, 2019; Miller, 2023), (ii) when the linear TWFE regression includes covariates (as in Firpo, 2007; Callaway and Sant'Anna, 2021), and (iii) when $y_{it}(0)$ is generated by a panel quantile regression.

Throughout this section, without any loss of generality, I will assume that there is no treatment anticipation (i.e., $\rho = 0$), that the policy is implemented starting from the second period onward, and that panel data are available. Let T denote the maximum number of available time periods, $q = 2$ the first time a policy is implemented, $r = \{2, \dots, T\}$ the first time a unit is treated, and n the sample size. I allow both n and T to vary in the following Monte Carlo exercises. I will compare the proposed estimators' performance in terms of average bias and root-mean-squared error (RMSE). For completeness, when considering the DGPs in which covariates play a role, I also consider the unconditional QTT parameter – which can be obtained using the identification result in Theorem 1 – where the distributional parallel trends and the copula invariance assumptions are assumed to hold unconditionally.

DGP 1: The first DGP considered is the following:

$$y_{it}(0) = \alpha_t + \eta_i + u_{it} \tag{2.4}$$

Where $y_{it}(0)$ denotes the untreated potential outcome, α_t are time fixed-effects, u_{it} represents the time-varying random shock, and η_i represents the time-invariant unobserved heterogeneity, which can be distributed differently across treated cohorts and never-treated units (i.e., I allow for selection on unobservables). Eq. (2.4) is similar to DGPs considered in Callaway et al. (2018); Callaway and Li (2019); Miller (2023).

Specifically, in *DGP 1*, I will set $\alpha_t = t$, $\eta | d_r = 1 \sim N(r, 1)$ with $r = \{2, \dots, T\}$, and $u_t \sim N(0, 1)$ in all periods. Implicitly assumed is that $T = R$, where R represents the last time period in which a unit can undergo the policy (i.e., there is a positive probability of being treated in every period starting from the second period onward). Since covariates play no role in this DGP, I set the probability of being first treated in period r, $\mathbb{P}(d_r = 1)$, to $1/T$. Then $y_{it}(0) \sim N(\alpha_t + r, 2)$. Exploiting the fact that for a generic random variable (R.V.) $X \sim N(\mu, \sigma^2)$, the τ-quantile is given by $F_X^{-1}(\tau) = \mu + \sigma\Phi^{-1}(\tau)$ – where $\Phi(\cdot)$ represents the standard normal cdf. Then the τth quantile at time t for units belonging to

[16] It is worth stressing that the rank invariance assumption can potentially hold in the following models since nothing is said about $y_t(r)|d_r = 1$.

cohort r is equal to $F^{-1}_{y_t(0)|d_r=1}(\tau) = (\alpha_t + r) + \sqrt{2}\Phi^{-1}(\tau)$.

DGP 2: The second DGP mimics the TWFE regression with covariates used by Callaway and Sant'Anna (2021) to study the finite-sample properties of the estimators of the group-time average treatment effects. Specifically, in contrast to *DGP 1*, it also allows for the selection on observable characteristics. I will consider the case where there is only one covariate $X \sim N(0,1)$ for simplicity. Since in this DGP, selection on observables is allowed, I set the probability to belong to cohort r to be a function of observable pre-treatment characteristics (as in Callaway and Sant'Anna, 2021):

$$\mathbb{P}(d_r = 1|X) = \frac{exp(X'\gamma_r)}{1 + \sum_r exp(X'\gamma_r)} \tag{2.5}$$

where $\gamma_r = 0.5r/T$. The following model generates the untreated potential outcome:

$$y_{it}(0) = \alpha_t + \eta_i + X_{it} + u_{it} \tag{2.6}$$

Compared to *DGP 1*, the only difference is that now I also allow selection on observable characteristics. I still assume $\alpha_t = t$, $\eta|d_r = 1 \sim N(r,1)$ with $r = \{2,\dots,T\}$, and $u_t \sim N(0,1)$. In this case, since $y_{it}(0) \sim N(\alpha_t + r, 3)$, the (population) τth quantile of $y_t(0)|d_r = 1$ is equal to $F^{-1}_{y_t(0)|d_r=1}(\tau) = (\alpha_t + r) + \sqrt{3}\Phi^{-1}(\tau)$.

In Appendix C, I present the results for *DGP 2*, but instead of assuming $\alpha_t = t$, I examine the estimator's performance in the presence of non-linear trends. Specifically, I assume a quadratic trend: $\alpha_t = t + t^2$.

DGP 3: The third DGP generates $y_{it}(0)$ by a panel quantile regression. Specifically, the following DGP is a re-adaptation of the DGP presented in Machado and Santos Silva (2019):

$$y_{it}(0) = \alpha_t + \eta_i + X_{it} + (1 + X_{it})\, u_{it} \tag{2.7}$$

where the probability of being treated is still given by (2.5), $u_t \sim N(0,1)$, and $\alpha_t = t$. Whereas now $\eta|d_r = 1 \sim Z_r$, $\gamma_r = \frac{r}{4T}$, $X_{it} = \frac{1}{8}\chi_{it}$, with $Z_r = r + T$, $T \sim \chi^2_{(1)}$ and $\chi_{it} \sim \chi^2_{(1)}$.

While in *DGP 1* and *DGP 2*, it was possible to obtain an analytical formula to compute the population quantiles of $y_t(0)|d_r = 1$, this is not the case for *DGP 3*. The distribution of $y_{it}(0)$ is difficult to retrieve. For this reason, I approximate the population quantiles with those obtained by simulating (2.7) using 1 million observations.

All DGPs presented in this section satisfy the assumptions needed for identification.[17] Specifically, in *DGP 1*, the distributional parallel trends and the copula invariance assumptions hold unconditionally. Whereas, in *DGP 2* and *DGP 3*, the distributional PT assumption holds conditionally on X, whereas the copula assumption still holds unconditionally.[18]

[17] Sufficient conditions for the distributional PT assumption and the copula invariance assumption to hold in linear models is that the error term is *iid* in each period. In the nonlinear DGP, this is complicated by the term $(1 + X_{it})u_{it}$. However, also in this case, one can prove that the distributional PT holds conditionally and the copula invariance holds unconditionally as long as X and u are independent of treatment assignment.

[18] The reason why the copula invariance still holds unconditionally is that none of the random variables in the right-hand side of the equation are distributed conditionally on X. As pointed out in the previous section, estimation may become demanding when the copula invariance holds conditionally.

In this subsection, I consider the following scenarios: i) $T = R = 4$ and ii) $T = R = 10$. I allow the sample size to vary in each setup, $n = \{100, 1000\}$. All results are based on $2,000$ Monte Carlo simulations, and I report the results for the .25, .50, and .75 quantiles. To save space only results for $F^{-1}_{y_2(0)|d_2=1}(\tau)$ are shown in the main text. Whereas in Appendix C, results obtained for scenario i) for the estimators of all the parameters when the quantile considered is .5 are presented. Based on the theoretical results obtained in Callaway and Li (2019) and Callaway and Sant'Anna (2021), the estimators are expected to perform well when n is large relative to R and T. What happens when T (and consequently R) grows is unclear, as the number of units in a given cohort can be very small.

Table 2.1: Monte Carlo Results for $F^{-1}_{y_2(0)|d_2=1,\rho=0}(\tau)$. Setup with $T = R = 4$.

			0.25			**0.50**			**0.75**		
	n	$F^{-1}_{y_t(0)\|d_r=1}$	Bias	Root MSE	$F^{-1}_{y_t(0)\|d_r=1}$	Bias	Root MSE	$F^{-1}_{y_t(0)\|d_r=1}$	Bias	Root MSE
DGP 1	**Unc**	100	3.046	0.112	0.525	4.000	0.097	0.489	4.954	0.122	0.509
DGP 2	**Unc**	100	2.832	0.331	0.695	4.000	0.341	0.684	5.168	0.391	0.714
	Cond	100	2.832	0.086	0.531	4.000	0.091	0.515	5.168	0.146	0.541
DGP 3	**Unc**	100	3.929	0.178	0.550	4.839	0.244	0.583	5.925	0.361	0.721
	Cond	100	3.929	0.114	0.571	4.839	0.203	0.590	5.925	0.335	0.730
DGP 1	**Unc**	1000	3.046	0.011	0.152	4.000	0.007	0.15	4.954	0.013	0.157
DGP 2	**Unc**	1000	2.832	0.258	0.314	4.000	0.25	0.304	5.168	0.257	0.315
	Cond	1000	2.832	0.018	0.154	4.000	0.011	0.147	5.168	0.018	0.158
DGP 3	**Unc**	1000	3.929	0.061	0.167	4.839	0.085	0.180	5.925	0.132	0.230
	Cond	1000	3.929	-0.005	0.163	4.839	0.040	0.169	5.925	0.101	0.218

Notes: This table reports the Monte Carlo results for the estimator of the parameter $F^{-1}_{y_2(0)|d_2=1,\rho=0}(\tau)$ in the setup with $T = R = 4$. Results are reported for the quantiles .25, .50, .75. Each Monte Carlo simulation uses $2,000$ bootstrap replications. Rows labeled 'DGP 1' report the results obtained for the DGP in Eq. (2.4), rows labeled 'DGP 2' present the results obtained for the DGP in Eq. (2.6), and rows labeled 'DGP 3' for the DGP in Eq. (2.7). Rows labeled 'UNC' use the estimator based on the unconditional distributional PT assumption that ignores covariates. Finally, $F^{-1}_{y_t(0)|d_r=1}$ represents the true population parameter, whereas 'Bias' and 'RMSE' stand for average (simulated) bias, and root-mean-squared, respectively.

Table 2.1 reports the simulation results for the estimators of $F^{-1}_{y_2(0)|d_2=1}(\tau)$ for scenario in which $T = R = 4$ (for all the three DGPs presented). What emerges is that, as expected, when the number of units per cohort is small ($n = 100$), the performance of the method presented is poor (especially true for the nonlinear DGP). As the number of units per cohort increases, the performance improves substantially in terms of bias and RMSE. Interestingly, the largest gain (in terms of bias reduction) is experienced when estimating the parameters of *DGP 3*. This is probably because, for n small, the proposed method 'struggles more' to retrieve the true counterfactual distribution when the DGP is non-linear (compared to a linear DGP), as few observations are present in each quantile. As n grows, however, the number of observations in each quantile increases, and the proposed method's performance increases.

Moreover, as the theory predicts, ignoring the role of covariates substantially reduces the estimator's performance, leading to unreliable inference.

Similar results are also found for the other cohort-time quantile treatment effects estimators. To save space, these results are omitted from the main text. Tables C.1, C.2 and C.3 in Appendix C, however, present the full set of results for the median – obtained when considering the scenario in which $T = 4$ – for *DGP 1*, *DGP 2* and *DGP 3*, respectively. All the other results are available upon request.

In Table C.4 in Appendix C, I present the results for the estimator of $F^{-1}_{y_2(0)|d_2=1}(\tau)$ in the scenario in which $T = R = 4$ for *DGP 2*, which now features a quadratic trend instead of a linear one.

The results indicate that the estimator's performance–both in terms of bias and RMSE–is nearly identical to the case where $\alpha_t = t$ shown in Table 2.1. This result is expected since, for Assumption 4 to hold, as long as the year fixed effects in untreated potential outcomes are independent of treatment assignment, any trend is accepted.

It is worth stressing that we would reach a similar conclusion to that just described above in the presence of non-linear trends if, instead of requiring the error term to be independent and identically distributed, we allowed the errors to be heteroskedastic (or within-cluster correlation to be present). As long as the structure of the heteroskedasticity/clustering is independent of treatment assignment, it does not lead to a violation of Assumption 4 or Assumption 5. Indeed, a sufficient condition for these two assumptions to be satisfied is that, even if clustering is present, the heteroskedasticity/within-cluster correlation remains unaffected by treatment assignment. Suppose, absurdly, that we know the skedastic function (e.g., $exp\{X\beta\}$ with $\beta \in \mathbb{R}$) and that $y_{it}(0)$ is generated according (2.4). A sufficient condition for the Distributional PT assumption to hold is that the skedastic function has a functional form that remains unchanged between treated groups and never-treated units. The Copula Invariance Assumption, on the other hand, requires that if autocorrelation or any type of correlation is present in the error term, the dependence structure (or correlation structure) must remain independent of treatment assignment.

Table 2.2: Monte Carlo Results for $F^{-1}_{y_2(0)|d_2=1,\rho=0}(\tau)$. Setup with $T = R = 10$.

			0.25			**0.50**			**0.75**		
		n	$F^{-1}_{y_t(0)\|d_r=1}$	Bias	Root MSE	$F^{-1}_{y_t(0)\|d_r=1}$	Bias	Root MSE	$F^{-1}_{y_t(0)\|d_r=1}$	Bias	Root MSE
DGP 1	**Unc**	100	3.046	0.047	0.858	4.000	0.073	0.864	4.954	0.137	0.895
DGP 2	**Unc**	100	2.832	0.114	1.087	4.000	0.196	1.082	5.168	0.306	1.091
	Cond	100	2.832	-0.02	0.927	4.000	0.085	0.911	5.168	0.173	0.918
DGP 3	**Unc**	100	3.935	0.091	0.925	4.861	0.209	0.996	5.971	0.377	1.337
	Cond	100	3.935	0.051	0.997	4.861	0.189	1.053	5.971	0.354	1.371
DGP 1	**Unc**	1000	3.046	0.013	0.245	4.000	0.005	0.232	4.954	0.013	0.251
DGP 2	**Unc**	1000	2.832	0.107	0.31	4.000	0.111	0.315	5.168	0.125	0.319
	Cond	1000	2.832	0.008	0.247	4.000	0.012	0.245	5.168	0.026	0.249
DGP 3	**Unc**	1000	3.935	0.044	0.261	4.861	0.069	0.274	5.971	0.100	0.327
	Cond	1000	3.935	-0.005	0.261	4.861	0.034	0.269	5.971	0.078	0.319

Notes: This table reports the Monte Carlo results for the estimator of the parameter $F^{-1}_{y_2(0)|d_2=1,\rho=0}(\tau)$ in the setup with $T = R = 10$. Results are reported for the quantiles .25, .50, .75. Each Monte Carlo simulation uses $2,000$ bootstrap replications. Rows labeled 'DGP 1' report the results obtained for the DGP in Eq. (2.4), rows labeled 'DGP 2' present the results obtained for the DGP in Eq. (2.6), and rows labeled 'DGP 3' for the DGP in Eq. (2.7). Rows labeled 'UNC' use the estimator based on the unconditional distributional PT assumption that ignores covariates. Finally, $F^{-1}_{y_t(0)|d_r=1}$ represents the true population parameter, whereas 'Bias' and 'RMSE' stand for average (simulated) bias, and root-mean-squared, respectively.

Regarding the scenario with $T = R = 10$, results are reported in Table 2.2 . Again, also in this case, I report the results for the estimators of $F^{-1}_{y_2(0)|d_2=1}(\tau)$ for the three DGPs. As with $T = R = 4$, when the number of units per cohort is small, the method's performance is poor, but as n increases, the method's performance improves substantially.

Interestingly, the average bias for the estimators of the parameters of all DGPs now appears to be smaller in magnitude than those shown in Table 2.1 (except for the $\tau = .75$ in the scenario with $n = 100$, where the magnitude is larger). However, for almost all cases, the RMSEs are larger than those shown in Table 2.1, so the parameters are now estimated imprecisely. The fact that the RMSEs are larger than before is probably because the number of units per cohort is much smaller.

Lastly, in the scenario with $T = R = 10$ and small n, the parameters of *DGP 2* and *DGP 3* could not be estimated in many simulations due to the low number of treated and/or never-treated units (results not shown).

So, overall, the performance of the estimators of the parameters of *DGP 2* and *DGP 3* does not necessarily improve compared to the scenario in which $T = R = 4$.

All the results reported above align with theoretical predictions and are consistent with the simulation results in Callaway et al. (2018); Callaway and Li (2019); Callaway and Sant'Anna (2021). The estimator's performance is relatively poor for small n but improves substantially as n increases. Specifically, all Monte Carlo results support the $\sqrt{n}$-consistency of the estimator for $F^{-1}_{y_t(0)|d_r=1,\rho}(\tau)$. Indeed, in both Table 2.1 and Table 2.2, the RMSE decreases approximately by a factor of $1/\sqrt{10}$ when the sample size increases from $n = 100$ to $n = 1000$, across all DGPs and regardless of the conditioning.

Lastly, in line with Callaway et al. (2018), I also found that, for all the DGPs, the method's power is relatively larger for the .50 quantile.

2.4.2 Monte-Carlo Simulations – violations of the main identifying assumptions

In this subsection, I analyze the performance of the proposed estimator in generating $F^{-1}_{y_t(0)|d_r=1}(\tau)$ when the untreated potential outcome follows a TWFE regression with covariates, similar to (2.6). However, in contrast to *DGP 2*, I now assess the impact of small deviations in:

- The Distributional Parallel Trends assumption (while the Copula Invariance assumption continues to hold), and
- The Copula Invariance assumption (while maintaining the Distributional Parallel Trends assumption).

As in the previous subsection, I assume no treatment anticipation, that the policy is implemented starting from the second period onward, and that panel data are available. I set $n = 1000$ and $T = R = 4$. Since covariates play a role, for completeness, I also evaluate the performance when considering the unconditional parameter. Further, as before, I assume that $X \sim N(0, 1)$, and the probability of belonging to cohort r follows Eq. (2.5), with $\gamma_r = 0.5r/T$.

DGP 4: In this DGP, I assess the performance of the proposed estimator in situations where Assumption 5 still holds unconditionally, as in *DGP 2*, but now I analyze the effects of small deviations from Assumption 4. To this end, I assume the untreated potential outcome follows the following TWFE regression:

$$y_{it}(0) = \alpha_{t,d} + \eta_i + X_{it} + u_{it} \tag{2.8}$$

where I still assume $\eta|d_r = 1 \sim N(r, 1)$ with $r = \{2, \ldots, 4\}$, and $u_t \sim N(0, 1)$. However, now $\alpha_{t,d} = t\,(1 + \bar{\varepsilon}d)$, where $d = 0$ for never-treated units and $d = 1$ for treated cohorts, and $\bar{\varepsilon}$

represents the degree of violation of Assumption 4 allowed. When $\bar{\varepsilon} = 0$, the Conditional Distributional Parallel Trends hold. When $\bar{\varepsilon} \neq 0$, Assumption 4 is violated. The larger $\bar{\varepsilon}d$ is, the greater the deviation considered.

By allowing $\alpha_{t,d}$ to vary, we introduce violations of the Conditional Distributional Parallel Trends between treated and never-treated units. Note that no distinction in violations between treated cohorts is considered, as this would complicate the analysis.

In this case, since $y_{it}(0) \sim N\left(t\left(1+\bar{\varepsilon}d\right)+r, 3\right)$, the (population) τth quantile of $y_t(0)|d_r = 1$ is equal to $F^{-1}_{y_t(0)|d_r=1}(\tau) = t\left(1+\bar{\varepsilon}d\right) + r + \sqrt{3}\Phi^{-1}(\tau)$.

Table 2.3: Monte Carlo Results, DGP 4, for $F^{-1}_{y_2(0)|d_2=1,\rho=0}(\tau)$ – Violation Cond. Distributional PT.

		0.25			**0.50**			**0.75**		
$\bar{\varepsilon}$		$F^{-1}_{y_t(0)\|d_r=1}$	Bias	Root MSE	$F^{-1}_{y_t(0)\|d_r=1}$	Bias	Root MSE	$F^{-1}_{y_t(0)\|d_r=1}$	Bias	Root MSE
0.00	**Unc**	2.832	0.257	0.314	4.000	0.253	0.307	5.168	0.253	0.311
	Cond	2.832	0.017	0.155	4.000	0.012	0.15	5.168	0.013	0.158
0.05	**Unc**	2.932	0.207	0.274	4.100	0.203	0.27	5.268	0.206	0.274
	Cond	2.932	-0.033	0.158	4.100	-0.037	0.158	5.268	-0.033	0.162
0.10	**Unc**	3.032	0.157	0.242	4.200	0.155	0.231	5.368	0.151	0.239
	Cond	3.032	-0.083	0.177	4.200	-0.086	0.166	5.368	-0.089	0.183
0.50	**Unc**	3.832	-0.239	0.297	5.000	-0.248	0.304	6.168	-0.252	0.312
	Cond	3.832	-0.48	0.504	5.000	-0.488	0.51	6.168	-0.492	0.516

Notes: This table reports the Monte Carlo results of the DGP in Eq. (2.8) for the estimator of the parameter $F^{-1}_{y_2(0)|d_2=1,\rho=0}(\tau)$ in the setup with $T = R = 4$ and violations of the Distributional PT assumption. Results are reported for the quantiles .25, .50, .75. Each Monte Carlo simulation uses $2,000$ bootstrap replications. Rows labeled 'UNC' use the estimator based on the unconditional distributional PT assumption that ignores covariates. Here, $\bar{\varepsilon}$ controls whether the Distributional PT is violated. When $\bar{\varepsilon} = 0$, Assumption 4 is not violated. When $\bar{\varepsilon} \neq 0$, Assumption 4 is violated. The larger is $\bar{\varepsilon}d$, the larger is the deviation considered. Finally, $F^{-1}_{y_t(0)|d_r=1}$ represents the true population parameter, whereas 'Bias' and 'RMSE' stand for average (simulated) bias, and root-mean-squared, respectively.

The results are shown in Table 2.3 and, where we consider $\bar{\varepsilon} \in \{0.00, 0.05, 0.10, 0.50\}$. For brevity, we report results only for the $F^{-1}_{y_2(0)|d_2=1,\rho=0}(\tau)$, though all other results are consistent with those in Table 2.3 and available upon request.

The results show that small violations of Assumption 4 lead to only minor increases in bias and RMSE. For instance, when focusing on the 50th quantile and considering the conditional parameter, the bias increases in magnitude from 0.012 (for $\bar{\varepsilon} = 0.00$) to 0.037 when $\bar{\varepsilon} = 0.05$ and to 0.086 when $\bar{\varepsilon} = 0.10$. The same applies to the RMSE. In contrast, for $\bar{\varepsilon} = 0.50$ (a large deviation from Assumption 4), the bias increases substantially. Similar results are observed for the 25th and 75th quantiles.

Furthermore, as demonstrated in the previous subsection, disregarding covariates significantly increases the estimator's bias and RMSE, except in cases where the violation of Assumption 4 is large. The poorer performance of the proposed estimator when conditioning on covariates in the presence of a large deviation from the Distributional PT assumption can be attributed to the substantial deviation in linear trends, which is not adequately captured by a single covariate. Since this covariate, by construction, has the same effect over time, it is unrelated to the trend in $y(0)$. Therefore, conditioning on X increases estimation noise within a given quantile.

DGP 5: In this last DGP, I assess the performance of the proposed estimator in scenarios where Assumption 4 still holds, but now I examine the effects of small deviations from unconditional Assumption 5, similar to the approach in DGP 2 in Callaway et al. (2018). Specifically, I assume that the untreated potential outcome follows the TWFE regression with covariates in (2.6) and that:

$$(\eta_i, u_{i,t}, u_{i,r-1}) \,|d_r = 1 \sim N(\boldsymbol{\mu}_r, V_r)$$

where $\boldsymbol{\mu}_r = [r, 0, 0]^T$ and

$$V_r = \begin{pmatrix} 1 & \rho_{r,u_t} & \rho_{r,u_{r-1}} \\ \rho_{r,u_t} & 1 & \rho_{u_t,u_{r-1}} \\ \rho_{r,u_{r-1}} & \rho_{u_t,u_{r-1}} & 1 \end{pmatrix} \quad r \in \{2,3,4\}, t \in \{1,\ldots,4\}$$

For simplicity, I assume V_r is symmetric; for example, $cov(u_t, u_{r-1}) = cov(u_{r-1}, u_t)$. Consequently, $(\Delta_{r-1,t} y(0), y_{r-1})$ follows a bivariate normal distribution with correlation parameter given by $\rho_{u_t,u_{r-1}} + \rho_{r,u_t} - \rho_{r,u_{r-1}} - 2$.

As noted by Callaway et al. (2018), when the bivariate distribution is normal, then the copula is Gaussian and the dependence parameter corresponds to the correlation coefficient.

To evaluate the estimator's performance under deviations from the Copula Invariance Assumption, I set $\rho_{u_t,u_{r-1}} = 0.5$ for all treated and never-treated units and define:

$$\rho_{r,u_t} = \begin{cases} \bar{\rho} d & \forall t \geq r \\ 0 & \forall t < r \end{cases}$$

where $d = 0$ for never-treated units and $d = 1$ for treated units (regardless of cohort), and $\bar{\rho}$ represents the degree of violation of Copula Invariance Assumption allowed. When $\bar{\rho} = 0$, the assumption holds, while $\bar{\rho} \neq 0$ indicates a violation. The larger $\bar{\rho}$ is, the greater the deviation considered.

Table 2.4 presents the results for $\bar{\rho} \in \{0.00, 0.05, 0.10, 0.50\}$. For brevity, we report results only for the $F^{-1}_{y_2(0)|d_2=1,\rho=0}(\tau)$, though all other results align with those in Table 2.4 and are available upon request.

Small violations of the Copula Invariance Assumption lead to minimal increases in the bias and RMSE is small. However, substantial violations (e.g., $\bar{\rho} = 0.50$) cause a pronounced increase in bias. For instance, considering the conditional parameter for the 0.25th quantile, the bias rises from 0.061 for $\bar{\rho} = 0.10$ to 0.203 for $\bar{\rho} = 0.50$. Similar patterns emerge for the 0.75th.

Conversely, the results for the .50th quantile are insensitive to deviations from the Copula Invariance Assumption. When allowing deviations from the Copula Invariance Assumption, only the variance of $y_{it}(0)$ is affected. However, $\Phi^{-1}(0.50) = 0$, thus we obtain $F^{-1}_{y_t(0)|d_r=1}(\tau) = (\alpha_t + r)$. These results align with those in Callaway et al. (2018).

Lastly, in nearly all the cases considered, ignoring covariates significantly increases estimator bias and RMSE. The only exception is for the 0.75th when $\bar{\rho} = 0.50$, where the bias is lower under the unconditional parameter–a puzzling result.

Table 2.4: Monte Carlo Results, DGP 5, for $F^{-1}_{y_2(0)|d_2=1,\rho=0}(\tau)$ – Violation Copula Invariance.

		0.25			0.50			0.75		
$\bar{\rho}$		$F^{-1}_{y_t(0)\|d_r=1}$	Bias	Root MSE	$F^{-1}_{y_t(0)\|d_r=1}$	Bias	Root MSE	$F^{-1}_{y_t(0)\|d_r=1}$	Bias	Root MSE
0.00	**Unc**	2.832	0.270	0.324	4.000	0.263	0.314	5.168	0.261	0.317
	Cond	2.832	0.023	0.153	4.000	0.017	0.146	5.168	0.015	0.16
0.05	**Unc**	2.812	0.290	0.341	4.000	0.263	0.314	5.188	0.241	0.302
	Cond	2.812	0.042	0.157	4.000	0.017	0.146	5.188	-0.005	0.159
0.10	**Unc**	2.793	0.308	0.357	4.000	0.263	0.314	5.207	0.222	0.287
	Cond	2.793	0.061	0.163	4.000	0.016	0.146	5.207	-0.024	0.161
0.50	**Unc**	2.651	0.449	0.484	4.000	0.263	0.314	5.349	0.080	0.197
	Cond	2.651	0.203	0.254	4.000	0.017	0.147	5.349	-0.166	0.23

Notes: This table reports the Monte Carlo results of the DGP in Eq. (2.6) for the estimator of the parameter $F^{-1}_{y_2(0)|d_2=1,\rho=0}(\tau)$ in the setup with $T = R = 4$ and violations of the Copula Invariance Assumption. Results are reported for the quantiles .25, .50, .75. Each Monte Carlo simulation uses $2,000$ bootstrap replications. Rows labeled 'UNC' use the estimator based on the unconditional distributional PT assumption that ignores covariates. Here, $\bar{\rho}$ controls whether the Copula Invariance is violated. When $\bar{\rho} = 0$, Assumption 5 is not violated. When $\bar{\rho} \neq 0$, Assumption 5 is violated. The larger is $\bar{\varepsilon}d$, the larger is the deviation considered. Finally, $F^{-1}_{y_t(0)|d_r=1}$ represents the true population parameter, whereas 'Bias' and 'RMSE' stand for average (simulated) bias, and root-mean-squared, respectively.

2.5 Discussion

The proposed method is only one of several possible ways to construct the counterfactual distribution. Estimating the counterfactual outcome is often challenging, as different estimators rely on different identifying assumptions[19] The performance of an estimator can vary depending on the context, so researchers should assess whether the proposed method is suitable for addressing their specific research question.

Several existing approaches identify the untreated potential outcome distribution for the treated group while allowing unobserved characteristics to vary across treated units. Using the intuition presented in this paper, these methods can be extended to the staggered treatment adoption setting.

The Changes-in-changes (CIC) model proposed by Athey and Imbens (2006) relates the outcome without intervention to the individual's group, time, and unobservable characteristics through a monotonic production function. While accounting for the selection on unobservables, the CIC model assumes that unobservable traits' distribution within a group is stable over time. Bonhomme and Sauder (2011), on the other hand, propose an estimator of the distributional treatment effects by requiring that the production function mapping groups, time, and covariates into outcome is additive.

Unlike Athey and Imbens (2006) and Bonhomme and Sauder (2011), the approach proposed in this paper does not restrict the functional form that relates groups, time, and covariates. Moreover, unlike Bonhomme and Sauder (2011), the proposed method does not require the unobservable time-varying component to be independent of treatment assignment. Nor does it assume that the time-varying unobservables are independent of time-invariant unobservables, conditional on treatment assignment.

The presented estimator requires only that the evolution of the untreated potential outcome be

[19] See Imbens and Wooldridge (2009) for a review of common assumptions in the causal inference literature for estimating counterfactual outcomes.

(conditionally) independent of treatment assignment while still allowing for serial correlation in the error term. Additionally, it permits the time-varying shock to be correlated with the individual (unobserved) heterogeneity. In this sense, the estimator proposed is more general than the other existing approaches, though this generality comes at the cost of an additional assumption regarding the missing dependence (copula) between the change in the untreated potential outcome and its pre-treatment level.

However, unlike the CIC model, the proposed approach is not scale-invariant. Additionally, in contrast to Bonhomme and Sauder (2011), the method does not allow for returns to unobserved skills to vary after introducing the policy.

It is also worth emphasizing that this estimator can complement existing methods for estimating the ATT when analyzing the impact of policies introduced in a staggered fashion. Given its similarity to Callaway and Sant'Anna (2021), I compare my estimator primarily to theirs, as the proposed approach extends their methodology.[20]

As discussed in previous sections, the proposed estimator complements existing estimators when heterogeneity along the outcome distribution is expetcted. While ATT is a widely used measure for evaluating policy effects, focusing on distributional treatment effects on the treated allows for broader applications, such as in equity assessments of policy interventions, making it relevant for policy makers. For example, in studying the effects of minimum wage policies on teen employment, as in Callaway and Sant'Anna (2021), the proposed estimator may provide deeper insights by capturing effects beyond the mean, particularly on the lower deciles of the income distribution.

While the proposed estimator offers advantages in settings where treatment effects are heterogeneous across the outcome distribution, it requires stronger identifying assumptions for consistency. Although Assumptions 1- 3 closely align with those in Callaway and Sant'Anna (2021), additional restrictions are needed beyond the Conditional Mean PT assumption to achieve point identification of $F_{y_t(0)|d_r=1}$. In particular, the estimator proposed is sensitive to strong violations of the Copula Invariance assumption (not required in Callaway and Sant'Anna, 2021) and may underperform, compared to the method proposed by Callaway and Sant'Anna (2021), when the PT assumption holds on average but not across the entire distribution.

Overall, when evaluating the effect of policy interventions, researchers should assess the plausibility of the identifying assumptions underlying this method relative to other available approaches.

2.6 Conclusion

In this paper, I provide a method to recover the whole distribution of the untreated potential outcome for the treated group in nonexperimental settings with staggered treatment adoption. To do so, I build on the idea behind the group-time average treatment effect estimator proposed by Callaway and Sant'Anna (2021) and generalize the existing quantile treatment effects on the treated estimator proposed by Callaway and Li (2019).

Once the causal parameters of interest are built, these parameters can be aggregated using aggregation schemes identical to those suggested by Callaway and Sant'Anna (2021) to highlight heterogeneity along specific dimensions (such as how treatment effects vary with the length of exposure to treatment).

[20] Interested readers may refer to Wooldridge (2021) for a comparison of the methods proposed by Callaway and Sant'Anna (2021), Sun and Abraham (2021), and Wooldridge (2021).

Further, I show that, in addition to the QTT, different approaches that anonymously summarize the quantiles of the distribution of the outcome of interest can be built within this framework once the entire counterfactual distribution is identified. Specifically, I consider tests for stochastic dominance rankings, which do not rely on the assumption of rank invariance, such as the QTT. Thus, this paper combines the literature on causal inference with the literature on inequality measures. Depending on the parameter of interest, I advocate that the researcher employ the most appropriate tests to perform statistical inference.

Identification is reached by extending the assumption of parallel trends, frequently invoked in empirical practice, to the entire distribution of the untreated potential outcome. Since this latter assumption is no longer sufficient to reach point identification (Fan and Yu, 2012), an additional assumption is required regarding the missing dependence (or copula) between the change in untreated potential outcome and its pretreatment level. To go back to point identification, following Callaway et al. (2018), I assume that the missing dependence is (conditionally) independent of the treatment assignment.

I show that, under the assumptions presented in this paper, the proposed method's performance is poor when the sample size is small, but this improves substantially as n increases. Furthermore, the proposed method performance remains almost unaffected even in cases with minor deviations from the main identifying assumptions. These findings align with the theoretical predictions established by Callaway et al. (2018); Callaway and Li (2019); Callaway and Sant'Anna (2021).

Bibliography

Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies*, 72(1):1–19.

Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497.

Bonhomme, S. and Sauder, U. (2011). Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling. *The Review of Economics and Statistics*, 93(2):479–494.

Borusyak, K., Jaravel, X., and Spiess, J. (2021). Revisiting event study designs: Robust and efficient estimation. *arXiv preprint:2108.12419*.

Butts, K. (2024). Difference-in-Differences Estimation with Spatial Spillovers. Papers 2105.03737, arXiv.org.

Callaway, B. and Li, T. (2019). Quantile treatment effects in difference in differences models with panel data. *Quantitative Economics*, 10(4):1579–1618.

Callaway, B., Li, T., and Oka, T. (2018). Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods. *Journal of Econometrics*, 206(2):395–413. Special issue on Advances in Econometric Theory: Essays in honor of Takeshi Amemiya.

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

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources*, 50(2):317–372.

Carneiro, P., Hansen, K. T., and Heckman, J. (2003). Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. IZA Discussion Papers 767, Institute of Labor Economics (IZA).

Chernozhukov, V., Fernandez-Val, I., and Kaji, T. (2016). Extremal quantile regression: An overview. *Handbook of Quantile Regression*.

Chernozhukov, V., Fernandez-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268.

Ciaccio, A. (2023). The impact of a cost-containment measure on the quality of regional health services in italy: a parametric and a non-parametric approach. Working Papers 2023: 24, Department of Economics, University of Venice "Ca' Foscari".

De Chaisemartin, C. and d'Haultfoeuille, X. (2022). Difference-in-differences estimators of intertemporal treatment effects. Technical report, National Bureau of Economic Research.

Fan, Y. and Yu, Z. (2012). Partial identification of distributional and quantile treatment effects in difference-in-differences models. *Economics Letters*, 115(3):511–515.

Fiorini, M., Lee, W., and Pfeifer, G. (2024). A simple approach to staggered difference-in-differences in the presence of spillovers. *CESifo Working Paper No. 11011*.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1):259–276.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, 66(5):1017–1098.

Heckman, J. J., Humphries, J. E., and Veramendi, G. (2016). Dynamic treatment effects. *Journal of Econometrics*, 191(2):276–292. Innovations in Measurement in Economics and Econometrics.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC press.

Li, X. and Lin, L. (2024). Identification of quantile treatment effects in difference-in-differences settings with staggered adoption. *Economics Letters*, 241:111792.

Linton, O., Maasoumi, E., and Whang, Y.-J. (2005). Consistent testing for stochastic dominance under general sampling schemes. *The Review of Economic Studies*, 72(3):735–765.

Maasoumi, E. (2003). *Parametric and Nonparametric Tests of Limited Domain and Ordered Hypotheses in Economics*, chapter 25, pages 538–556. John Wiley & Sons, Ltd.

Maasoumi, E. and Wang, L. (2019). The gender gap between earnings distributions. *Journal of Political Economy*, 127(5):2438–2504.

Machado, J. A. and Santos Silva, J. (2019). Quantiles via moments. *Journal of Econometrics*, 213(1):145–173.

MacKinnon, J. G., Nielsen, M. Ø., and Webb, M. D. (2023). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics*, 232(2):272–299.

MacKinnon, J. G. and Webb, M. D. (2018). The wild bootstrap for few (treated) clusters. *The Econometrics Journal*, 21(2):114–135.

Manski, C. F. (1994). Analog estimation of econometric models. *Handbook of econometrics*, 4:2559–2582.

Manski, C. F. and Pepper, J. V. (2013). Deterrence and the death penalty: Partial identification analysis using repeated cross sections. *Journal of Quantitative Criminology*, 29:123–141.

Melly, B. and Santangelo, G. (2015). The changes-in-changes model with covariates. *W.P.*

Miller, B. I. (2023). Doubly-robust quantile treatment effect estimation. *W.P.*

Nelsen, R. (2006). *An Introduction to Copulas*. Springer.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512.

Robins, J. M. (1987). Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". *Computers & Mathematics with Applications*, 14(9):923–945.

Roth, J., Sant'Anna, P. H., Bilinski, A., and Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*.

Sant'Anna, P. H. and Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231.

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

Wooldridge, J. M. (2021). Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators. *Available at SSRN 3906345*.

Xu, R. (2025). Difference-in-differences with interference.

Appendix A – Proofs

Identification

Identification without covariates

In this subsection, following Callaway and Li (2019) I will prove the result obtained in Theorem 1. To prove this theorem, I will use two results of Sklar's Theorem: Lemma A.1. and Lemma A.2. in Appendix A in Callaway and Li (2019). To save space, I refer the reader to their paper.

Proof of Theorem 1. For notational convenience, I will assume that there is no treatment anticipation (i.e., $\rho = 0$) and abbreviate the joint pdf at time t of the change in untreated potential outcome and the pre-treatment untreated potential outcome for treated group r as $f_{t|d_r=1}(\cdot,\cdot) = f_{y_t(0)-y_{r-\rho-1}(0),y_{r-\rho-1}(0)|d_r=1}$. Similarly, this will be, instead, the same pdf for the never-treated $f_{t|C=1}(\cdot,\cdot) = f_{y_t(0)-y_{r-\rho-1}(0),y_{r-\rho-1}(0)|C=1}$. Further, I will denote the copula pdfs between the change in untreated potential outcome and the pre-treatment untreated potential outcome as $c_{t|d_r=1}(\cdot,\cdot) = c_{y_t(0)-y_{r-\rho-1}(0),y_{r-\rho-1}(0)|d_g=1}$ where $g = \{r, 0\}$. Assuming also that $\Delta_{[r-1,t]}y(0)$ has support in $\Delta\mathcal{Y}$ and $y_{t-1}(0)$ in $\mathcal{Y}$, then:

$$
\begin{aligned}
F_{y_t(0)|d_r=1} &= \mathbb{P}\left(y_t(0) \leq y | d_r = 1\right) \\
&= \mathbb{P}\left(\Delta_{[r-1,t]}y(0) + y_{r-1}(0) \leq y | d_r = 1\right) \\
&= \mathbb{E}\left[\mathbb{1}\left\{\Delta_{[r-1,t]}y(0) \leq y - y_{r-1}(0) | d_r = 1\right\}\right] \\
&= \int_{\mathcal{Y}}\int_{\Delta\mathcal{Y}} \mathbb{1}\left\{\delta \leq y - y'\right\} f_{t|d_r=1}\left(\delta, y'\right) d\delta dy' \\
&= \int_{\mathcal{Y}}\int_{\Delta\mathcal{Y}} \mathbb{1}\left\{\delta \leq y - y'\right\} c_{t|d_r=1}\left(F_{\Delta_{[r-1,t]}y(0)|d_r=1}, F_{y_{r-1}(0)|d_r=1}\right) \\
&\quad \times f_{\Delta_{[r-1,t]}y(0)|d_r=1}(\delta) f_{y_{r-1}(0)|d_r=1}(y') d\delta dy' \qquad (A.1) \\
&= \int_{\mathcal{Y}}\int_{\Delta\mathcal{Y}} \mathbb{1}\left\{\delta \leq y - y'\right\} c_{t|C=1}\left(F_{\Delta_{[r-1,t]}y(0)|d_r=1}, F_{y_{r-1}(0)|d_r=1}\right) \\
&\quad \times f_{\Delta_{[r-1,t]}y(0)|d_r=1}(\delta) f_{y_{r-1}(0)|d_r=1}(y') d\delta dy' \qquad (A.2) \\
&= \int_{\mathcal{Y}}\int_{\Delta\mathcal{Y}} \mathbb{1}\left\{\delta \leq y - y'\right\} f_{t|C=1}(F^{-1}_{\Delta_{[r-1,t]}y(0)|C=1}\left(F_{\Delta_{[r-1,t]}y(0)|d_r=1}(\delta)\right), \\
&\quad F^{-1}_{y_{r-1}(0)|C=1}\left(F_{y_{r-1}(0)|d_r=1}(y')\right)) \\
&\quad \times \frac{f_{\Delta_{[r-1,t]}y(0)|d_r=1}(\delta)}{f_{\Delta_{[r-1,t]}y(0)|C=1}\left(F_{\Delta_{[r-1,t]}y(0)|d_r=1}(\delta)\right)} \\
&\quad \times \frac{f_{y_{r-1}(0)|d_r=1}(y')}{f_{y_{r-1}(0)|C=1}\left(F^{-1}_{y_{r-1}(0)|C=1}\left(F_{y_{r-1}(0)|d_r=1}(y')\right)\right)} d\delta dy' \qquad (A.3)
\end{aligned}
$$

Equation (A.1) exploits Lemma A.1 in Callaway and Li (2019) to write the joint distribution using a copula pdf. Equation (A.2) exploits Assumption 5 to retrieve the missing dependence between the change in untreated potential outcome and the pre-treatment value of untreated potential outcome for treated group r. Lastly, equation (A.3) employs Lemma A.2 in Callaway and Li (2019) to rewrite the copula pdf as the joint distribution for the never-treated group.

Let us now make the following change of variables to simplify computations. Specifically, let us

denote with:

$$u = F^{-1}_{\Delta_{[r-1,t]}y(0)|C=1}\left(F_{\Delta_{[r-1,t]}y(0)|d_r=1}(\delta)\right), \quad v = F^{-1}_{y_{r-1}(0)|C=1}\left(F_{y_{r-1}(0)|d_r=1}(y')\right)$$

The above notation then implies the following equalities:

1. $\delta = F^{-1}_{\Delta_{[r-1,t]}y(0)|d_r=1}\left(F_{\Delta_{[r-1,t]}y(0)|C=1}(u)\right)$
2. $y' = F^{-1}_{y_{r-1}(0)|d_r=1}\left(F_{y_{r-1}(0)|C=1}(v)\right)$
3. $\frac{d\delta}{du} = \frac{f_{\Delta_{[r-1,t]}y(0)|C=1}(u)}{f_{\Delta_{[r-1,t]}y(0)|d_r=1}\left(F^{-1}_{\Delta_{[r-1,t]}y(0)|d_r=1}\left(F_{\Delta_{[r-1,t]}y(0)|C=1}(u)\right)\right)}$
4. $\frac{dy'}{dv} = \frac{f_{y_{r-1}(0)|C=1}(v)}{f_{y_{r-1}(0)|d_r=1}\left(F^{-1}_{y_{r-1}(0)|d_r=1}\left(F_{y_{r-1}(0)|C=1}(v)\right)\right)}$

If we plug (1)-(4) in equation (A.3), we obtain the following equalities:

$$\begin{aligned}
&= \int_{\mathcal{Y}}\int_{\Delta\mathcal{Y}} \mathbb{1}(F^{-1}_{\Delta_{[r-1,t]}y(0)|d_r=1}\left(F_{\Delta_{[r-1,t]}y(0)|C=1}(u)\right) \leq y - \\
&\quad F^{-1}_{y_{r-1}(0)|d_r=1}\left(F_{y_{r-1}(0)|C=1}(v)\right)) \times f_{t|C=1}(u,v)dudv \quad (A.4)\\
&= \mathbb{E}\Big[\mathbb{1}\left(F^{-1}_{\Delta_{[r-1,t]}y(0)|d_r=1}\left(F_{\Delta_{[r-1,t]}y(0)|C=1}\left(\Delta_{[r-1,t]}y(0)\right)\right) \leq y - F^{-1}_{y_{r-1}(0)|d_r=1}\times\right.\\
&\quad \left.\left(F_{y_{r-1}(0)|C=1}\left(y_{r-1}(0)\right)\right)\Big] \,|C=1\Big] \quad (A.5)\\
&= \mathbb{E}\Big[\mathbb{1}\left(F^{-1}_{\Delta_{[r-1,t]}y(0)|C=1}\left(F_{\Delta_{[r-1,t]}y(0)|C=1}\left(\Delta_{[r-1,t]}y(0)\right)\right) \leq y - F^{-1}_{y_{r-1}(0)|d_r=1}\times\right.\\
&\quad \left.\left(F_{y_{r-1}(0)|C=1}\left(y_{r-1}(0)\right)\right)\Big] \,|C=1\Big] \quad (A.6)\\
&= \mathbb{E}\left[\mathbb{1}\left(\Delta_{[r-1,t]}y(0)\right) \leq y - F^{-1}_{y_{r-1}(0)|d_r=1}\left(F_{y_{r-1}(0)|C=1}\left(y_{r-1}(0)\right)\right)|C=1\right] \quad (A.7)
\end{aligned}$$

Where equation (A.4) comes from substituting previous identities in equation (A.3); equation (A.5) comes from the definition of $\mathbb{E}(\cdot)$; equation (A.6) replaces the unknown distribution $F^{-1}_{\Delta_{[r-1,t]}y(0)|d_r=1}$ with the distribution of the change in untreated potential outcome for the never-treated group thanks to Assumption 4 holding unconditionally, and lastly equation (A.7) proves the result as each of these distributions of untreated potential outcomes is identified by their sample counterparts. This proves identification of $F_{y_t(0)|d_r=1}$.

□

Alternative Proof of Theorem 1. An alternative and more direct approach to proving the identification of $F_{y_t(0)|d_r=1}$ is to exploit Assumption 6 and Sklar's Theorem, as done in Callaway et al. (2018).

To see this, recall that for every $y \in supp\left(y_t(0)|d_r=1\right)$, we can express $F_{y_t(0)|d_r=1}$:

$$\begin{aligned}
F_{y_t(0)|d_r=1} &= \mathbb{P}\left(y_{it}(0) \leq y | d_r = 1\right)\\
&= \mathbb{P}\left(\Delta_{[r-1,t]}y_i(0) + y_{i,r-1}(0) \leq y | d_r = 1\right)
\end{aligned} \tag{A.8}$$

Under the continuity assumption, we can write, for units treated in cohort r:

$$\Delta_{[r-1,t]}y_i(0) = F^{-1}_{\Delta_{[r-1,t]}y(0)|d_r=1}\left(u_i^r\right) \quad \text{and} \quad y_{i,r-1}(0) = F^{-1}_{y_{r-1}(0)|d_r=1}\left(v_i^r\right) \tag{A.9}$$

where $u_i^r \equiv F_{\Delta_{[r-1,t]}y(0)|d_r=1}\left(\Delta_{[r-1,t]}y_i(0)\right)$ and $v_i^r \equiv F_{y_{i,r-1}(0)|d_r=1}\left(y_{i,r-1}(0)|d_r=1\right)$.

Similarly, for the never-treated:

$$\Delta_{[r-1,t]}y_i(0) = F^{-1}_{\Delta_{[r-1,t]}y(0)|C=1}\left(u_i^0\right) \quad \text{and} \quad y_{i,r-1}(0) = F^{-1}_{y_{r-1}(0)|C=1}\left(v_i^0\right) \tag{A.10}$$

where $u_i^0 \equiv F_{\Delta_{[r-1,t]}y(0)|C=1}\left(\Delta_{[r-1,t]}y_i(0)\right)$ and $v_i^0 \equiv F_{y_{i,r-1}(0)|C=1}\left(y_{i,r-1}(0)|C=1\right)$.

Thus, substituting these into (A.8), we obtain:

$$F_{y_t(0)|d_r=1} = \mathbb{P}\left(F^{-1}_{\Delta_{[r-1,t]}y(0)|d_r=1}\left(u_i^r\right) + F^{-1}_{y_{r-1}(0)|d_r=1}\left(v_i^r\right) \leq y|d_r=1\right) \tag{A.11}$$

For $d_r = 1$, the joint distribution $F_{\left(u_i^r,v_i^r\right)|d_r=1}$ is unknown. However, by Sklar's Theorem, we express it in terms of conditional copula $C_{\Delta y_t(0),y_{t-1}(0)|X,d_r=1}$. By the unconditional version of assumption Assumption 5, we replace $C_{\Delta y_t(0),y_{t-1}(0)|d_r=1}$ with $C_{\Delta y_t(0),y_{t-1}(0)|C=1}$.

Substituting this into (A.11), we obtain:

$$F_{y_t(0)|d_r=1} = \mathbb{P}\left(F^{-1}_{\Delta_{[r-1,t]}y(0)|d_r=1}\left(u_i^0\right) + F^{-1}_{y_{r-1}(0)|d_r=1}\left(v_i^0\right) \leq y|C=1\right)$$

While $F^{-1}_{y_{r-1}(0)|d_r=1}$ is observable from the data, $F^{-1}_{\Delta_{[r-1,t]}y_{(0)}|d_r=1}$ is not. Using Assumption 4, we replace the unknown distribution $F^{-1}_{\Delta_{[r-1,t]}y(0)|d_r=1}$ with the distribution of the change in untreated potential outcome for the never-treated group. Further, using (A.10) and the definitions of u_i^0 and v_i^0, we obtain:

$$F_{y_t(0)|d_r=1} = \mathbb{P}\left(\Delta_{[r-1,t]}y_i(0) + F^{-1}_{y_{r-1}(0)|d_r=1}\left(F_{y_{r-1}(0)|C=1}\left(y_{i,r-1}(0)\right)\right) \leq y|C=1\right)$$

which completes the proof. □

Identification with covariates

In the first part of this subsection I will prove the identification result shown Proposition 1; in the second part of this subsection, instead, I will prove Proposition 2.

Proof of Proposition 1. As pointed out in Section 2.2, all the results obtained in Theorem 1 are still valid. The only part that changes is equation (A.6), which used an unconditional version of Assumption 4 to reach identification of $F_{\Delta y(0)|d_r=1}$. Now, this object is identified by the reweighted distribution in Eq. (2.2). To show that the results obtained in Theorem 1 are still valid what we need to prove is that

$F_{\Delta y(0)|d_r=1} = F^p_{\Delta y(0)|d_r=1}$. To prove it, let us exploit the definition of $F_{\Delta y(0)|d_r=1}$:

$$
\begin{aligned}
F_{\Delta y(0)|d_r=1} &= \mathbb{P}(\Delta y_t(0) \leq \delta | d_r = 1) \\
&= \frac{\mathbb{P}(\Delta y_t(0) \leq \delta, d_r = 1)}{p_r} \quad (A.12) \\
&= \mathbb{E}\left(\frac{\mathbb{P}(\Delta y_t(0) \leq \delta, d_r = 1|X)}{p_r}\right) \\
&= \mathbb{E}\left(\frac{P_r(X)}{p_r}\mathbb{P}(\Delta y_t(0) \leq \delta | d_r = 1, X)\right) \\
&= \mathbb{E}\left(\frac{P_r(X)}{p_r}\mathbb{P}(\Delta y_t(0) \leq \delta | X, C = 1)\right) \quad (A.13) \\
&= \mathbb{E}\left(\frac{P_r(X)}{p_r}\mathbb{E}\left[C\mathbb{1}\{\Delta y \leq \delta\}|X, C = 1\right]\right) \quad (A.14) \\
&= \mathbb{E}\left(\frac{P_r(X)}{p_r(1-p_r(X))}\mathbb{E}\left[C\mathbb{1}\{\Delta y_t \leq \delta\}|X\right]\right) \\
&= \mathbb{E}\left(\frac{Cp_r(X)}{p_r(1-p_r(X))}\mathbb{1}\{\Delta y_t \leq \delta\}\right) \quad (A.11)
\end{aligned}
$$

where p_r denotes the probability of being treated in cohort r, and $p_r(X) = \mathbb{P}\left(d_r = 1|X, d_r + C = 1\right)$ denotes the generalized propensity score, as defined in Assumption 7.

In Equation (A.12) I exploited the definition of conditional probability; Equation (A.13) holds since Assumption 4 holds; in Equation (A.14) I exploited the definition of probability, and then I multiplied by C (this holds since $\mathbb{E}(\cdot)$ is conditionally on $C = 1$). Further, by conditioning on $C = 1$, we can rewrite the potential outcome as the observed outcome. The last equality exploits the Law of Iterated Expectations and concludes the proofs.

□

Proof of Proposition 2. The proof of Proposition 2 follows directly from Theorem 1, where now all the steps hold after conditioning on covariates.

□

Appendix B – Additional Results for Repeated Cross Sections

In this section, I extend the identification results of this paper to the setting where repeated cross sections are available. To this end, I will follow Callaway et al. (2018) and Callaway and Sant'Anna (2021).

Specifically, I assume that for each cross-sectional unit, the researcher has access to $(y, d_r, \ldots, d_T, C, t, X)$ where $r = q, \ldots, T$ and $t = 1, \ldots, T$ is the period in which unit i is observed. Additionally, if we denote with S_t a dummy variable taking value 1 if an observation is observed in period t and zero otherwise.

Now, I assume that a random sample is available in each period.

Assumption 1. *Conditional on period t, $(y, d_r, \ldots, d_T, C, X)$ are cross sectionally independent and identically distributed for all $t \in \{1, \ldots, T\}$, where $(d_r, \ldots, d_T, C, X)$ is invariant to t.*

Assumption 1 is identical to that done in Callaway and Sant'Anna (2021) and means that the pooled cross section is composed of independent draws from the following mixture distribution:

$$F_M(y, d_r, \ldots, d_T, c, t, x) = \sum_{t=1}^{T} \mathbb{P}(S_t = 1) F_{y,d_r,\ldots,d_T,C,t,X}(y, d_r, \ldots, d_T, c, x|t)$$

This assumption is also similar to the assumption in Abadie (2005) and Sant'Anna and Zhao (2020), and excludes the possibility of compositional changes over time.

For simplicity, suppose that covariates do not play a role in identification. Further, assume without loss of generality that there is no treatment anticipation(i.e., $\rho = 0$). To extend the main identification result from Theorem 1, we impose an additional requirement on the data-generating process. Even if Assumptions 1, 1, 6, and an unconditional version of Assumptions 3- 5 hold, the difference $y_{it} - y_{i,r-1}$ is not observed for the same unit in a pooled cross section.

To achieve point identification, following Callaway et al. (2018), I impose a *rank invariance* assumption on potential outcomes over time. This assumption allows us to recover the (unobserved) outcome at period t by exploiting the rank of the outcome observed in $r - 1$. It assumes that units preserve their relative position in the distribution of y over time.

Although rank invariance is a strong assumption and is often rejected in empirical data (as discussed in the main text), it is necessary to recover the unknown distribution at time t.

Corollary 1. *Suppose we have access to repeated cross sectional data, specifically $\{(y, d_{ir}, C_i)\}_{i=1}^{n^s}$ for period $s \in \{r-1, t\}$ where $r = q, \ldots, T$, $t = 1, \ldots, T$, and n^s denotes the sample size of the cross section. Suppose further that Assumptions 1, 6, B1, and the unconditional version of Assumptions 3-5 hold.*

If the copula of $(y_{i,r-1}(0), y_{i,t}(0)|C = 1)$ satisfies rank invariance, then for every $(u, v) \in [0, 1]^2$

$$C_{y_{i,r-1}(0),y_{i,t}(0)|C=1}(u, v) = min\{u, v\}$$

Thus, for $y \in supp\,(y_{i,t}(0)|d_r = 1)$, we obtain

$$F_{y_t(0)|d_r=1}(y) = \mathbb{P}\{\tilde{\Delta}_{[r-1,t]}y(0) + F^{-1}_{y_{r-1}(0)|d_r=1}\left(F_{y_{r-1(0)|C=1}}(y_{r-1}(0))\right) \leq y|C = 1\}$$

where $\tilde{\Delta}_{[r-1,t]}y(0) \equiv F^{-1}_{y_t(0)|C=1}\left(F_{y_{r-1(0)|C=1}}(y_{r-1}(0))\right) - y_{r-1}$.

As noted in Callaway et al. (2018), the rank invariance assumption on the copula of $(y_{i,r-1}(0), y_{i,t}(0)|C = 1)$ neither implies nor is implied by Assumption 5.

Proof of Corollary 1. Under Assumptions 1, 6, B1, and the unconditional version of Assumptions 3-5, the result from Theorem 1 holds:

$$F_{y_t(0)|d_r=1} = \mathbb{P}\{\Delta_{[r-1,t]}y(0) + F^{-1}_{y_{r-1}(0)|d_r=1}\left(F_{y_{r-1}(0)|C=1}\left(y_{r-1}(0)\right)\right) \leq y|C=1\}$$

for $y \in supp\left(y_{i,t}(0)|d_r=1\right)$.

However, we are working with repeated cross sections, $\Delta_{[r-1,t]}y(0)$ cannot be identified from the observed outcome for the never-treated, as in Theorem 1. To solve for this issue, we impose the rank invariance assumption, which yields:

$$F_{y_t(0)|C=1}\left(y_t\left(0\right)\right) = F_{y_{r-1}(0)|C=1}\left(y_{r-1}\left(0\right)\right)$$

since both $F_{y_t(0)|C=1}\left(\cdot\right)$ and $F_{y_{r-1}(0)|C=1}\left(\cdot\right)$ are identifiable from observed outcomes, we can express $\Delta_{[r-1,t]}y(0)$ for units with $C=1$ as

$$\tilde{\Delta}_{[r-1,t]}y(0) \equiv F^{-1}_{y_t(0)|C=1}\left(F_{y_{r-1(0)|C=1}}\left(y_{r-1}(0)\right)\right) - y_{r-1}$$

which completes the proof. □

This result extends naturally to cases where covariates are included. Specifically, while the identification result in Proposition 2 can be generalized by ensuring that all steps hold conditional on X, extending Proposition 1 requires a generalized definition of the generalized propensity score, as outlined in Appendix B of Callaway and Sant'Anna (2021).

Appendix C – Additional Simulation Results

Table C.1: Monte Carlo Results, DGP 1, $\tau = .5$. Setup with $T = R = 4$.

		$n = 100$		$n = 1,000$	
	$F^{-1}_{y_t(0)\|d_r=1}$	Bias	Root MSE	Bias	Root MSE
$F^{-1}_{y_2(0)\|d_2=1}$	4	0.097	0.489	0.007	0.15
$F^{-1}_{y_3(0)\|d_2=1}$	5	0.089	0.491	0.009	0.148
$F^{-1}_{y_4(0)\|d_2=1}$	6	0.097	0.492	0.008	0.148
$F^{-1}_{y_3(0)\|d_3=1}$	6	0.087	0.451	0.013	0.133
$F^{-1}_{y_4(0)\|d_3=1}$	7	0.095	0.451	0.012	0.133
$F^{-1}_{y_4(0)\|d_4=1}$	8	0.101	0.416	0.015	0.126

Notes: This table reports the Monte Carlo results of the DGP in Eq. (2.4) in the setup with $T = R = 4$ for the quantile $\tau = .5$. Each Monte Carlo simulation uses $2,000$ bootstrap replications. Rows labeled 'UNC' use the estimator based on the unconditional distributional PT assumption that ignores covariates. Finally, $F^{-1}_{y_t(0)|d_r=1}$ represents the true population parameter, whereas 'Bias' and 'RMSE' stand for average (simulated) bias, and root-mean-squared, respectively.

Table C.2: Monte Carlo Results, DGP 2, $\tau = .5$. Setup with $T = R = 4$.

			$n = 100$		$n = 1,000$	
		$F^{-1}_{y_t(0)\|d_r=1}$	Bias	Root MSE	Bias	Root MSE
Unc	$F^{-1}_{y_2(0)\|d_2=1}$	4	0.341	0.684	0.25	0.304
	$F^{-1}_{y_3(0)\|d_2=1}$	5	0.344	0.686	0.248	0.303
	$F^{-1}_{y_4(0)\|d_2=1}$	6	0.341	0.689	0.248	0.302
	$F^{-1}_{y_3(0)\|d_3=1}$	6	0.469	0.723	0.371	0.41
	$F^{-1}_{y_4(0)\|d_3=1}$	7	0.47	0.728	0.372	0.41
	$F^{-1}_{y_4(0)\|d_4=1}$	8	0.571	0.794	0.479	0.509
Cond	$F^{-1}_{y_2(0)\|d_2=1}$	4	0.091	0.515	0.011	0.147
	$F^{-1}_{y_3(0)\|d_2=1}$	5	0.099	0.522	0.007	0.147
	$F^{-1}_{y_4(0)\|d_2=1}$	6	0.091	0.519	0.005	0.145
	$F^{-1}_{y_3(0)\|d_3=1}$	6	0.099	0.508	0.014	0.156
	$F^{-1}_{y_4(0)\|d_3=1}$	7	0.098	0.515	0.012	0.151
	$F^{-1}_{y_4(0)\|d_4=1}$	8	0.09	0.507	0.007	0.153

Notes: This table reports the Monte Carlo results for all the parameters of the DGP in Eq. (2.6) in the setup with $T = R = 4$ for the quantile $\tau = .5$. Each Monte Carlo simulation uses $2,000$ bootstrap replications. Rows labeled 'UNC' use the estimator based on the unconditional distributional PT assumption that ignores covariates. Finally, $F^{-1}_{y_t(0)|d_r=1}$ represents the true population parameter, whereas 'Bias' and 'RMSE' stand for average (simulated) bias, and root-mean-squared, respectively.

Table C.3: Monte Carlo Results, DGP 3, $\tau = .5$. Setup with $T = R = 4$.

			$n = 100$		$n = 1,000$	
		$F^{-1}_{y_t(0)\|d_r=1}$	Bias	Root MSE	Bias	Root MSE
Unc	$F^{-1}_{y_2(0)\|d_2=1}$	4.839	0.244	0.583	0.085	0.18
	$F^{-1}_{y_3(0)\|d_2=1}$	5.841	0.239	0.571	0.083	0.179
	$F^{-1}_{y_4(0)\|d_2=1}$	6.867	0.2	0.55	0.061	0.171
	$F^{-1}_{y_3(0)\|d_3=1}$	6.837	0.239	0.561	0.092	0.178
	$F^{-1}_{y_4(0)\|d_3=1}$	7.869	0.206	0.544	0.064	0.163
	$F^{-1}_{y_4(0)\|d_4=1}$	8.867	0.206	0.527	0.065	0.162
Cond	$F^{-1}_{y_2(0)\|d_2=1}$	4.839	0.203	0.59	0.04	0.169
	$F^{-1}_{y_3(0)\|d_2=1}$	5.841	0.195	0.576	0.036	0.168
	$F^{-1}_{y_4(0)\|d_2=1}$	6.867	0.16	0.56	0.015	0.167
	$F^{-1}_{y_3(0)\|d_3=1}$	6.837	0.191	0.573	0.041	0.168
	$F^{-1}_{y_4(0)\|d_3=1}$	7.869	0.16	0.565	0.012	0.163
	$F^{-1}_{y_4(0)\|d_4=1}$	8.867	0.157	0.545	0.017	0.158

Notes: This table reports the Monte Carlo results for all the parameters of the DGP in Eq. (2.7) in the setup with $T = R = 4$ for the quantile $\tau = .5$. Each Monte Carlo simulation uses $2,000$ bootstrap replications. Rows labeled 'UNC' use the estimator based on the unconditional distributional PT assumption that ignores covariates. Finally, $F^{-1}_{y_t(0)|d_r=1}$ represents the true population parameter, whereas 'Bias' and 'RMSE' stand for average (simulated) bias, and root-mean-squared, respectively.

Table C.4: Monte Carlo Results, DGP2, for $F^{-1}_{y_2(0)|d_2=1,\rho=0}(\tau)$. Setup with $T = R = 4$ and non-linear trends.

		0.25			**0.50**			**0.75**		
		$F^{-1}_{y_t(0)\|d_r=1}$	Bias	Root MSE	$F^{-1}_{y_t(0)\|d_r=1}$	Bias	Root MSE	$F^{-1}_{y_t(0)\|d_r=1}$	Bias	Root MSE
$n = 100$	**Unc**	6.832	0.331	0.684	8	0.344	0.672	9.168	0.392	0.715
	Cond	6.832	0.082	0.526	8	0.09	0.508	9.168	0.152	0.553
$n = 1000$	**Unc**	6.832	0.256	0.315	8	0.252	0.305	9.168	0.25	0.31
	Cond	6.832	0.017	0.158	8	0.013	0.149	9.168	0.011	0.156

Notes: This table reports the Monte Carlo results for the estimator of the parameter $F^{-1}_{y_2(0)|d_2=1,\rho=0}(\tau)$ of the DGP in Eq. (2.6) in the setup with $T = R = 4$ and with a quadratic trend (i.e., $\alpha_t = t + t^2$). Results are reported for the quantiles .25, .50, .75. Each Monte Carlo simulation uses $2,000$ bootstrap replications. Rows labeled 'UNC' use the estimator based on the unconditional distributional PT assumption that ignores covariates. Finally, $F^{-1}_{y_t(0)|d_r=1}$ represents the true population parameter, whereas 'Bias' and 'RMSE' stand for average (simulated) bias, and root-mean-squared, respectively.

The Camorra Kills Even without Guns. The Impact of the Land of Fires on Health Outcomes

Andrea Ciaccio[1], Domenico Depalo[2], Irene Mammi[1],

[1]Department of Economics, Ca' Foscari University of Venice

[1]Bank of Italy, Labor Market Department, Italy

Abstract

An increasing body of scientific literature shows that exposure to environmental contamination due to waste mismanagement negatively affects health outcomes. In this paper, we estimate the causal effect of the Land of Fires (Terra dei Fuochi or TdF) – an area of 90 municipalities between the provinces of Naples and Caserta (South of Italy), where, starting from the end of the 1980s, illegal dumping and burning of (toxic) waste has occurred because of organized crime – on health outcomes. Estimating the average treatment effect via a doubly robust estimand, we find an excess of cancer-specific mortality in the Land of Fires compared to the rest of the region. This result holds for both sexes and across different specifications.

JEL codes: C14, C21, C23, I18, J38
Keywords: Land of Fires, Illegal Waste Dumping, Environmental Exposure, Mortality from Cancer, Doubly-robust method
Acknowledgments: I am grateful to Davide Raggi, Federico Belotti, Paolo Li Donni, and Mario Padula for their invaluable advice, insights, and detailed feedback on previous drafts of this paper.

3.1 Introduction

Uncontrolled and illegal waste disposal poses significant health risks, as it may contaminate groundwater, soil, and air, thus creating severe conditions for the population residing in areas affected by such phenomena. Besides, the illegal burning of hazardous waste can directly negatively affect the health of nearby communities (WHO, 2024). For these reasons, preventing and implementing actions aimed at defeating and counteracting these criminal activities is at the top of policymakers' agenda.

From the beginning of the 2000s, numerous studies were conducted to understand whether individuals living in Campania – a region from Southern Italy – were exposed to increased incidence and mortality from specific malignancies due to exposure to hazardous waste. It is in Campania and, more specifically, in a particular area of 90 municipalities between the provinces of Naples and Caserta that, since the end of the 1980s, illegal dumping and burning of hazardous waste have occurred because of *Camorra*. This area is known as the Land of Fires (or, in Italian, *Terra dei Fuochi*). In 2013, the regional and central government, the provinces of Naples and Caserta, and the mayors of 57 municipalities (later becoming 90 by 2015) of the provinces of Naples and Caserta signed an agreement (*Patto Terra dei Fuochi*) aimed at implementing numerous measures to counteract the uncontrolled and illegal dumping of urban and toxic waste. Notwithstanding, hazardous waste is still being dumped and burned. Furthermore, the phenomenon seems to have increased substantially over the years, thus calling for more drastic policy interventions from the central government.

Despite various studies having been conducted by public health authorities and researchers to identify a link between exposure to dumping and burning hazardous waste and increased incidence and mortality from cancer and leukemia (compared to the rest of the region and the rest of Italy), these studies only provide an estimate of this relationship without estimating the causal effect (Alberti, 2022).

This paper aims to bridge the gap in the existing literature by estimating the causal effect of the Land of Fires on health outcomes using data from different administrative sources. Specifically, we consider as health outcomes cancer-specific mortality and mortality for all causes of death. In addition, compared to the existing literature, we use a much larger time span, from 2003 to 2019, which is crucial as it could take years to detect any adverse effect.

To estimate the causal effect of the Land of Fires, we exploit causal inference methods that allow us to retrieve estimates of the average treatment effect (ATE) and average treatment effect on the treated (ATT) by controlling for observed demographics and socio-economic differences that may correlate with treatment status. Failing to account for observed differences in socio-demographic characteristics may bias the results. Controlling for a large set of covariates, as we do, is paramount as it allows us to compare two otherwise similar municipalities, but because one is situated in TdF and the other not.

Specifically, we employ one commonly used technique: the *inverse probability weighted regression adjustment* (or shortly IPWRA) estimator. We also re-run the analysis using the *propensity-score matching* as a robustness check. The propensity-score matching imputes the missing potential outcome for each unit by averaging the outcomes of similar subjects who received the opposite treatment – where the similarity between treated and untreated units is assessed using an estimate of the propensity score. By contrast, the IPWRA estimator applies probability weights to derive outcome-regression parameters that adjust for the selection problem. For each unit, we observe only one of the two potential outcomes (Holland, 1986). These adjusted parameters are then used to compute averages of predicted outcomes for treated and control units. The differences between these averages represent the estimate of the ATE.

As the existing literature on causal inference suggests, doubly robust estimators must be preferred to propensity-score matching as they are more robust against model misspecification.

Our findings indicate that individuals residing in the Land of Fires die more from cancer than in the rest of the region and also compared to a counterfactual scenario where treated municipalities would have been unaffected by the treatment. Our results align perfectly with what ISS (2020) reports. In addition, results are robust to different sensitivity analyses and robustness checks performed.

Overall, the fact that people living in Terra dei Fuochi experience a higher cancer mortality rate than the rest of the region supports the hypothesis that environmental and air contamination exposure harms individuals' health. This calls for more drastic policy interventions from the central government and local authorities to defeat and counteract these criminal activities.

This paper is structured as follows. Section 3.2 outlines the institutional context. Section 3.3 reviews the existing epidemiological studies on the Land of Fires. In Section 3.4, we describe how the dataset used for the analysis was constructed and the indicators employed. Moreover, different descriptive statistics are also provided in this section. Section 3.5 reports the methods used to estimate the causal effect of Terra dei Fuochi. In Section 3.6, the main results obtained are presented alongside different robustness checks. A discussion of these results is offered in light of the existing literature. Lastly, Section 3.7 concludes.

3.2 Institutional Setting

Campania is a region in the Southern part of Italy. Its population accounted for almost 9.6% of the overall population in 2019, making it Italy's third most populated region – following Lombardy and Lazio. Besides being the most populated region in the South, Campania has Italy's youngest population. Notwithstanding, it is characterized by substantial within-region variability, with some municipalities of the region's hinterland experiencing an incidence of people aged 75 or older well above the average of the rest of Italy (Istat, 2020). In addition, Campania represents the region with the highest population density in Italy, half of which lives in the province of Naples. It is in this province that Italy's first six municipalities for population density are located (Istat, 2020; Svimez, 2022).

Despite being one of the most populated regions, Campania is one of the poorest regions in Italy. Specifically, it features the highest number (in absolute terms) of poor households. Moreover, it is characterized by the highest deprivation index in the South of Italy.[1] Further, the region is also characterized by substantial income inequalities between municipalities, with some municipalities at the top of the taxable income distribution (*reddito IRPEF*). In contrast, others reside at the bottom of this distribution (NVVIP, 2013).

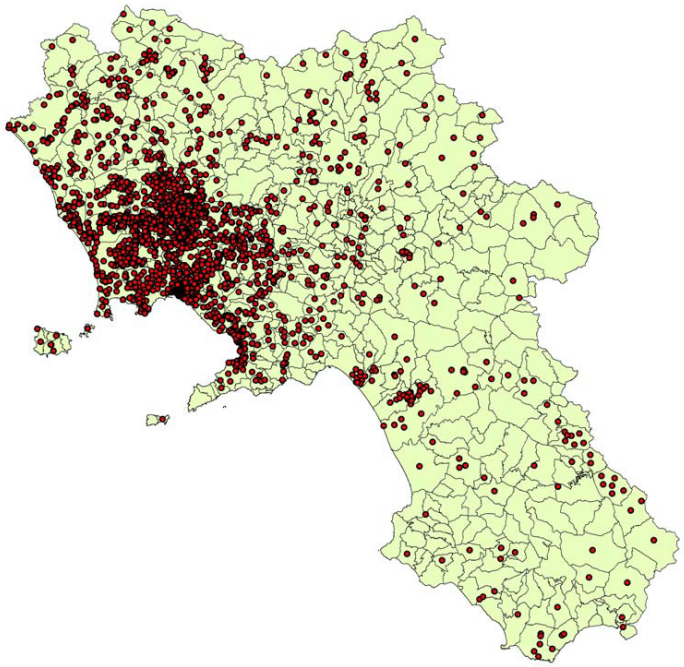
In this context, one of Italy's most vital criminal organizations, *Camorra*, found a fertile ground for its business. Indeed, in the late 1980s, Camorra took over the management of – first urban and then industrial – waste, transforming it into a flourishing business – estimated to amount to 600 million Euros per year (Tortora, 1989; Saviano, 2006; Legambiente, 2013; Mazza et al., 2018).

Specifically, according to different investigations, these urban and industrial wastes – including, among others, asbestos, heavy metals, and battery acids – were disposed of in all the region's landfills.

[1]The deprivation index expresses the level of relative social disadvantage through a combination of specific characteristics of a given population of interest. The various factors are chosen to represent the different conditions of economic disadvantage.

Most importantly, the waste not dumped in the landfills was buried illegally underground or in unauthorized landfills. Such a phenomenon has occurred especially in an area of approximately 800 square kilometers between the province of Naples and the southwestern part of Caserta, turning these territories into open-air landfills, which was later internationally recognized as the "Land of Fires" (*Terra dei Fuochi* or, shortly, *TdF*),[2] because of waste combustion (Legambiente, 2003; Ministero dell'Interno, 2018).[3]

Figure 3.1: Illegal and potentially at risk waste dumping sites in Campania region in 2005.



Notes: Source ARPAC (2005).

In Figure 3.1, it is reported, as of March 2005, the spatial distribution of illegal waste (dumping and burning) sites identified by the study carried out by the Regional Agency for Environmental Protection of Campania, ARPAC (ARPAC, 2005).[4] It can be easily seen from the picture that the most considerable fraction of illegal dumping sites is situated in the Land of Fires.

Numerous scientific studies have shown that exposure to uncontrolled disposal of hazardous waste and environmental pollution adversely affects the health of the population (Jarup et al., 2002; Garcia-Perez et al., 2013; WHO, 2024).

Given the existence of possible adverse health effects, researchers and public health authorities, starting from the early 2000s, conducted numerous studies to understand whether living in TdF or, more generally, in Campania, might be associated with worse health outcomes (Trinca et al., 2001;

[2]For the remainder of the paper, I will use the words Land of Fires, Terra dei Fuochi or TdF indiscriminately.

[3]For further details, please refer to the Regional Agency for Environmental Protection of Campania (ARPAC): https://www.arpacampania.it/terra-dei-fuochi.

[4]Unfortunately, a more updated version of this map is not publicly available at the time of the writing. However, the number of illegal dumping sites that are potentially at risk increased from 1,505 spread across all the region (ARPAC, 2005) to 2,767 identified only in the provinces of Naples and Caserta between 2008 and 2015 (ISS, 2020).

ARPAC, 2005; Altavista et al., 2004; Senior and Mazza, 2004).

While findings about some specific outcomes (e.g., survival after surgical treatment of lung cancer) are mixed (Rocco et al., 2016; Mazza et al., 2018), there exists a unanimous consensus that living in TdF is associated with increased incidence and mortality from some specific types of cancer and leukemia, frequent congenital anomalies when compared to both regional and national data (Senior and Mazza, 2004; Pirastu et al., 2013; Mazza et al., 2015; Fusco et al., 2020; ISS, 2020; Cafieri and Feoli, 2023; Fazzo et al., 2023). Results can be explained by the fact that the soil and groundwater of these areas are highly contaminated and no longer suitable for agriculture. This is due to the high presence of toxic waste as well as dioxin and polychlorinated biphenyls (PCB) levels well above the safety threshold (Altavista et al., 2004; ISS, 2020; Cafieri and Feoli, 2023).[5]

3.2.1 Patto Terra dei Fuochi

The first studies pointing out adverse health effects for people residing in the areas mentioned above were conducted in the early 2000s. Moreover, by 2008, numerous illegal dumping sites were already sized by local authorities (De Crescenzo, 2021). Notwithstanding, for almost two decades, "reactions of regional and national health authorities to these findings were generally aimed at reassurance of local communities and did not explicitly recognize a risk to public health" (Alberti, 2022, p.3).

The Land of Fires, indeed, only caught the attention of policymakers in 2013: this year, local, regional, and central authorities first signed an agreement, *Patto per la Terra dei Fuochi* (d.M.I. 26/11/2012). This was the first time in three decades that it was officially acknowledged that the uncontrolled and illegal disposal of waste and its burning might have detrimental health effects.

This pact aimed to implement numerous measures to counteract unlawful activities, including increased supervision of urban and extra-urban roads. Moreover, frequent inspections of manufacturing businesses recognized as generally providing the raw material for fires were dictated (e.g., car workshops, tire dealers, shops, and companies related to the textile industry).

A two-tier system was enforced: on the one hand, the region promoted cooperation between those municipalities adhering to the agreement and allotted 5 million Euros to finance projects whose primary goal was to prevent unlawful waste disposal. On the other hand, those same municipalities were required to identify and remove all waste disposed of on public grounds. In addition, an open-source database was published on the Prometeo website, aiming at geolocalizing all the sites where illegal disposal and burning of waste occurred.

3.2.2 DL 136/2013 "Terra dei Fuochi"

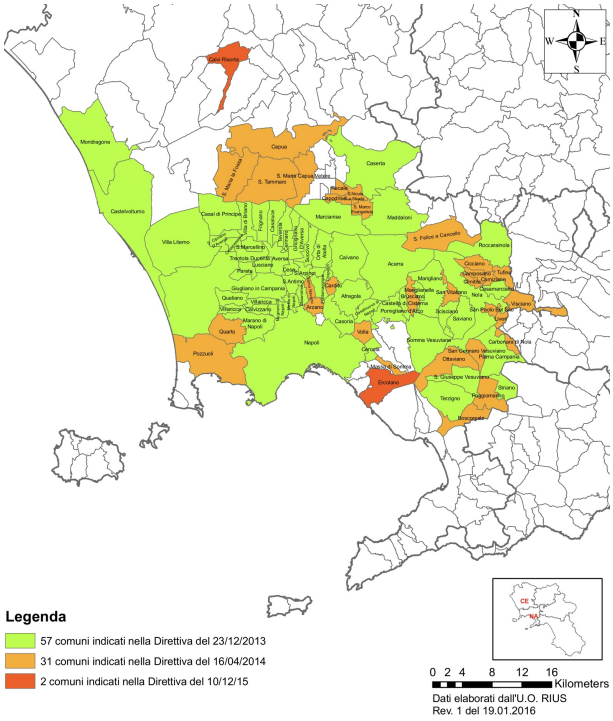
In 2013, the Central Government issued a Decree, "Terra dei Fuochi" (DL 136/2013), containing a series of dispositions to deal with the severe environmental emergency in the affected areas.

This decree (later converted into Law 6 of 2014) dictated the mapping of agricultural lands likely to have been affected by the contamination. Where the analyses found a significant presence of toxic materials, the decree required confiscating and forbidding their use for agricultural purposes. Simultaneously, it dictated the decontamination of highly polluted lands.

[5] For a thorough review of the studies conducted over the years and an overview of the results found, please refer to Mazza et al. (2018) and Alberti (2022).

Based on the National Institute of Health (ISS) proposal, the central government also required the activation of screening and prevention programs to monitor the population's health status, allocating 25 million Euros annually to reach this goal.

Figure 3.2: Municipalities belonging to the Land of Fires.



Notes: Source ARPAC. For further details, please visit https://www.arpacampania.it/documents/20195/152582/TDF_tutti+i+comuni+delle+due+Direttive+rev+1+def.pdf/a64eb63a-104a-4238-9081-f995dc6008bd.

Lastly, the decree dictated the active prosecution of illegal waste disposal.

Nowadays, the Central Government officially recognizes the Land of Fires as composed of 90 municipalities, depicted in Figure 3.2. Specifically, in green are depicted municipalities officially recognized in 2013, in orange are those that joined the agreement Terra dei Fuochi later in 2014, and in red are the last two municipalities joining the agreement in 2015.

Although ten years have passed since the agreement was first signed, the government has never carried out any land decontamination or implemented other measures to defeat these crimes.

Despite the numerous attempts from the municipalities and the region to counteract these phenomena, illegal dumping and burning waste continue.

The activation of cancer registries during the last decade increased the epidemiological surveillance of the population residing in these territories and Campania more generally.

3.3 Background Literature

Numerous epidemiological and medical studies document that exposure to environmental pollution, waste incinerator emissions, hazardous waste disposal, as well as living near landfills are associated with adverse health outcomes – such as increased incidence of asthma and cancer (Jarup et al., 2002; Garcia-Perez et al., 2013; Fazzo et al., 2017; WHO, 2024).

Given the severe environmental emergency of municipalities situated in the Land of Fires, starting from the early 2000s, numerous researchers, as well as the National Institute of Health (ISS) have analyzed whether the uncontrolled waste dumping and burning had any adverse effect (Trinca et al., 2001; Senior and Mazza, 2004; Mazza et al., 2015; Rocco et al., 2016; Fusco et al., 2020; ISS, 2020; Cafieri and Feoli, 2023; Fazzo et al., 2023).

These studies can be broadly categorized into two strands. The first strand focused on understanding whether the widespread dumping and burning of waste was associated with an increased risk of biological and environmental contamination. Other studies, instead, have analyzed whether there is an association between hazardous waste and the increased incidence and mortality for some specific malignancies (such as cancer) known to be related to environmental pollution.

Rivezzi et al. (2013) analyzed the milk of 94 breastfeeding women aged 19-32 living in the affected provinces and found that all samples contained an average level of dioxins of 16.6 pg TEQ/g of fat (well above the safety threshold identified by the European Commission, which is 3 pg TEQ/g fat). These results align with data on blood and serum samples of almost 860 participants and data on the milk of 52 women in the study conducted by De Felip et al. (2014). While heavy-metals concentrations in serum and blood samples align with values detected in the rest of Europe, dioxin concentrations in breastfeeding mothers' milk were positively associated with living in contaminated areas. Maselli et al. (2010) collected 200 frogs from contaminated and uncontaminated areas of the Campania Region. The authors found that frogs collected from potentially contaminated areas reported higher levels of DNA damage than in the rest of the region.

Contrasting evidence found by independent researchers, the University of Naples Federico II and the Zooprophylactic Institute in Portici report that soil and groundwater contamination from dioxins or heavy metals is restricted mainly to specific micro-areas near urban or industrial locations, affecting less than 5% of the Land of Fires (Bencivelli, 2016). ARPAC (2017) finds similar results on the levels of contamination.

The second strand of the literature focused on proving the adverse health effects. Trinca et al. (2001) and Lauria and Spinelli (2004), in a report conducted for ISS, found that childhood mortality in the province of Caserta was higher near landfills.

Altavista et al. (2004) studying cause-specific mortality ratios of Giugliano, Quagliano, and Villaricca – three municipalities situated in the province of Naples – found no difference in the cancer-specific mortality rate between Quagliano and Villaricca compared to the rest of the region. In contrast, the cancer-specific mortality ratio experienced by people residing in Giugliano was well above the regional average.

Rocco et al. (2016), instead, analyzing the 5-year survival rates of 439 patients diagnosed with lung cancer operated between November 2004 and April 2013 at the Division of Thoracic Surgery of the National Cancer Institute of Naples did not find any statistically significant difference between patients residing in TdF and those coming from other areas of the region. By contrast, accounting for the socioeconomic deprivation index, Fusco et al. (2020) report an excess of cancer-specific mortality in TdF. Similarly, Cafieri and Feoli (2023) – using data on pollution and mortality coming from different administrative sources – find descriptive evidence that mortality is higher in the affected areas.[6]

Lastly, some studies have attempted to directly link information on the type of waste disposal and

[6]Other studies exist on this topic, conducted by researchers or public health authorities. For an overview, please refer to Mazza et al. (2015), Mazza et al. (2018), Alberti (2022).

the geolocalization of illegal dumping to health outcomes. For instance, a report conducted by the ISS considered pathologies for which the evidence of an association with hazardous waste was defined as "sufficient" or "limited" in the scientific literature (ISS, 2020; Fazzo et al., 2023). Specifically, they analyzed whether there were systematic differences between 38 municipalities belonging to TdF (those belonging to Procura di Napoli 38) and the rest of the region, proving excess mortality and incidence for all types of cancer in these areas. These results seem to be driven by excess mortality and incidence of liver and bladder cancer in both genders and of breast cancer in women, as well as an increase in mortality for leukemias for people aged 0-19. Lastly, via regression analysis, they find a positive correlation between cancer-specific mortality and an indicator of waste risk at the municipal level. The Municipal Risk Index was built by leveraging spatial information on the waste type and quantity, waste disposal typology, and population close to the illegal landfills (Fazzo et al., 2020).

However, as pointed out by Alberti (2022), all these studies fail to identify the causal effect of living in the Land of Fires on health outcomes, stemming from three main issues. First, the slow-developing nature of many types of cancers makes it difficult to identify a direct link between exposure to waste and the onset of these malignancies. Second, the population affected is large and characterized by substantial heterogeneity regarding other possible confounders of cancer (e.g., different socioeconomic conditions). Third, and most importantly, there is limited information about the type, amount, and distribution of toxic substances disposed of. Altogether, this suggests that additional work is needed to estimate the causal effect of TdF on health outcomes.

3.4 Data

To provide evidence on the causal impact of the Land of Fires on health-related outcomes, we built a unique dataset gathering information from different administrative sources.[7]

Data on area (in km^2), mortality (cause-specific and all-causes of death), population, and municipalities' statistical classifications come from the National Institute of Statistics (ISTAT). Data on municipalities belonging to the Land of Fires were gathered from the ministerial directives of 23/12/2013, 16/04/2014, and 10/12/15. Lastly, data on taxable household income come from the Ministry of Economics and Finance.

Regarding sample selection, data have already been aggregated at the municipality level and include information on all 550 municipalities in the Campania region. On one side, this level of granularity is sufficient to capture regional variation between municipalities and, more importantly, to understand whether municipalities in TdF differ from the rest of the region, both in terms of mortality and sociodemographic characteristics. On the other side, having data already aggregated at the municipality level limits subsequent analysis. For instance, if individual-level data had been available for research purposes, one could have analyzed whether individuals residing closer to illegal dumping sites die more from specific pathologies than those living farther away. Further, we do not have access to information on the nature and the quantity of toxic material disposed of on the grounds, nor the exact year the illegal dumping started in a specific site. Thus, it is impossible to investigate whether there are heterogeneous effects between municipalities, years of exposure, or the type of toxic waste buried in the

[7] Unfortunately, we can only consider mortality as a health outcome. For privacy reasons, data on the incidence of specific malignancies are unavailable at the time of writing.

ground.[8]

As for the time span, except for the taxable household income (for which we have information starting from 2001), we have information from 1982 to 2019. However, for cause-specific mortality, the data are coded according to the 10^{th} revision of the ICD (ICD-9-CM) from 2003 only. To avoid jumps in the time series due to variations in the classification of diseases, we opted to restrict our analysis to 2003 to 2019. Even though we cannot evaluate whether pre-existing trends in cancer-specific mortality already existed in these areas before the illegal dumping phenomenon began, having access to a time span of seventeen years should be sufficient to capture variations between municipalities over time and detect region-specific trends.

Further, as done by previous studies (see, for instance, ISS, 2020), we decided to exclude the five municipalities representing the provincial capitals of the region (or in Italian *capoluoghi di provincia*) from the analysis, as these latter differ significantly in terms of sociodemographic characteristics (as also pointed out by Svimez, 2022).[9] These municipalities are Avellino, Benevento, Caserta, Napoli, and Salerno.

Overall, we have $9,265$ municipality-year information.

3.4.1 Mortality (all causes of death) and cause-specific mortality

To evaluate the impact of the Land of Fires, we exploited as the main outcome of interest mortality for those pathologies for which the evidence of an association with hazardous waste was defined as "sufficient" or "limited" in the scientific literature, as done in the existing literature on TdF (Fusco et al., 2020; ISS, 2020; Fazzo et al., 2023).

Specifically, we considered *cancer-specific mortality*, for which a correlation is proved with illegal waste disposal. In addition, we also analyzed the impact of TdF on *mortality for all causes of death* as done by Fusco et al. (2020) and Musmeci et al. (2015). In contrast to these latter studies, we did not adjust for the deprivation index. This choice is due to the recent debate on the appropriateness of such an indicator to represent the socio-economic diversity of the population of interest (Pasetto et al., 2017).

Data on mortality come from the National Institute of Statistics. They are administrative data collected yearly for all the residing population in Italy via the "Survey on Deaths and Causes of Death", regardless of where the death occurred (Italian territory or abroad). More specifically, these mortality data are conventionally based on a single cause of death (or underlying cause). The definition of underlying cause employed by ISTAT is the following:

- "the disease or injury that started that chain of morbid events that leads directly to the death", or
- "the set of circumstances of the accident or violence that caused the fatal traumatic injury".

To prevent misclassification, ISTAT uses IRIS, a software designed to automatically code multiple causes of death and determine the primary cause. The software aims to standardize coding practices in accordance with WHO recommendations (Istat, 2018).

While data on all causes of death are available by municipality of residence, year, sex, and age group, data on cause-specific mortality are not available by age group for privacy concerns. Besides,

[8] While individual-level data are not released due to privacy concerns of the legislator; information on the type and quantity of pollutants disposed of in the grounds is complex to obtain given the unregulated and illegal nature of the disposal.

[9] These descriptive statistics are not shown to save space but are available upon request.

the cause is not released in cases in which the frequency of deaths by municipality and sex is less than 3. For instance, suppose that in the town of Acerra, there were two female deaths in 2010. Then, for privacy concerns, ISTAT does not release the cause associated with these two deaths.

To study the impact of TdF, we decided to focus on mortality by the municipality of residence due to the slow-developing nature of cancer-specific malignancies.

Moreover, to account for the residing population in a specific municipality and thus compare municipalities with similar population sizes, we decided to consider two indicators commonly used in the literature. The first indicator is the *mortality rate per ten thousand inhabitants*. Specifically, to define the mortality rate, we used the definition commonly used by ISTAT:

$$\text{mortality rate}_{magt} = \frac{death_{magt}}{population_{magt}} \times 10,000$$

where $death_{magt}$ represents the absolute number of deaths occurring in municipality m in age group a and sex g at time t, and $population_{magt}$ represents the residing population at time t in municipality m in the age group a and sex g.

The second indicator, instead, is the *mortality ratio (MR)*. To compute this latter indicator, we follow instead the definition used by ISS (2020):

$$\text{mortality ratio}_{magt} = \frac{death_{magt}}{\text{expected deaths}_{magt}} \times 100$$

where $\text{expected deaths}_{magt} = population_{magt} \times mr_{agt}$ and $mr_{agt} = \frac{death_{agt}}{population_{agt}}$ denotes the regional mortality rate.

The mortality ratio compares the number of deaths observed in the population under analysis with the expected number of deaths it would have experienced had the population of interest faced the mortality levels of the reference population (in this case, the region). The MR, therefore, expresses, as a percentage, the excess or deficiency of mortality of the population under analysis compared to the reference population.

The ratio takes a value of 100 if the population of interest experiences the same mortality as the reference population, greater than 100 if there is excess mortality in the population under study, and lower than 100 if there is a deficiency.

Regarding population data used to compute the denominators of the two indicators, we used those published by the National Institute of Statistics on DemoIstat, an open-source online database containing different characteristics of the Italian-residing population. Data on the resident population are available from 1980 to 2019 for all ages, sexes, and citizenship (Italians vs. non-Italians).

It should be emphasized that while mortality data are collected annually for the entire resident population, population data are derived from intercensal estimates by age, sex, and citizenship as of January 1st. To ensure a high level of accuracy, ISTAT periodically reviews published estimates. Furthermore, although estimation errors may exist in these data, as long as there are no systematic differences between population estimates for municipalities in TdF and the rest of the region, this should not affect our analysis.

3.4.2 Socio-demographic characteristics

A set of socioeconomic variables was also obtained to understand whether affected municipalities differed from the rest of the region regarding population size, density, distribution by age and gender, and other socioeconomic characteristics. Failing to account for observed differences may lead to a spurious correlation between TdF and health outcomes.

Specifically, we obtained data on *taxable household income* from the Ministry of Economics and Finance. These data have already been aggregated at the municipal level and are publicly available.[10] To account for different populations sizes, we divided the taxable household income by the respective population size, thus generating a sort of income per capita.

It is worth highlighting that data on taxable household income are derived from individual income tax return files. As a result, underreporting is highly likely. However, if there are no systematic differences between the municipalities in TdF and the rest of the region, this should not threaten our analysis.

All other characteristics were obtained from the National Institute of Statistics. These include: *population density per* km^2, *population size*, and *population distribution by age*, *altitude zone*, *altitude (in meters)*, whether it is a *coastal* or *insular* municipality, the *coastal area*, and the *level of urbanization*.

Specifically, *altitude zone* is a categorical variable that classifies the altitude zone of the municipality into mountain, hill, and plain.

Altitude represents the height (in meters) above the sea level of the municipality. *Coastal area* is a dummy variable taking value 1 if the municipality is located on the coast or has at least 50% of the surface at a distance from the sea of less than 10 km.

level of urbanization is a categorical variable that takes a value of 1 if the municipality has a high population density; a value of 2 if it is a small town; and a value of 3 if it is a rural area. [11]

Lastly, since *population size* enters the denominator of our two health indicators, in the following analysis, we use pre-treatment values of this variable and its related distribution by age. As explained in the following sections, rather than using current values, we use the average of these variables computed over the period from 1980 to 1989 (pre-treatment period). For instance, in Section 3.6, when estimating the average treatment effect of the TdF on the mortality rate at time t, population size and its distribution by age will be included as covariates. However, to avoid any 'endogeneity' issues, we replace their values observed at time t with their sample means computed using years from 1980 to 1989.

3.4.3 Descriptive Statistics

Before estimating the causal effect of the Land of Fires on our health outcomes, we performed a battery of descriptive statistics to understand whether living in TdF was positively associated with increased mortality. We report some of them hereafter. It is worth stressing that we excluded "capoluoghi di provincia" from all the subsequent analyses.

Table 3.1 reports descriptive statistics on mortality for all causes of death (indicated by *a.c.*), cancer-specific mortality, and population at the municipal level split by gender and treatment status. Specifi-

[10] For further details, please refer to https://www1.finanze.gov.it/finanze/analisi_stat/public/index.php?search_class[0]=cCOMUNE&opendata=yes.

[11] These data are publicly available at the following link: https://www.istat.it/classificazione/principali-statistiche-geografiche-sui-comuni/.

Table 3.1: Descriptive statistics – aggregated.

		No Land of Fires						Land of Fires					
	Sex	**N**	**Mean**	**SD**	**Median**	**Min**	**Max**	**N**	**Mean**	**SD**	**Median**	**Min**	**Max**
Population	Males	7769	2737.663	4456.109	1235	117	43766	1496	10694.06	9579.588	7655	762	59341
Deaths (a.c.)	Males	7769	26.48449	40.3137	14	0	442	1496	78.87767	70.75632	55	4	400
m-rate (a.c.)	Males	7769	114.9391	48.43193	105.9964	0	500	1496	75.32404	18.49651	73.23096	20.71108	160.9195
m-ratio (a.c.)	Males	7769	132.3827	55.48247	122.3361	0	542.5284	1496	86.7123	20.91096	84.5814	24.54403	184.0708
Deaths (cancer)	Males	7769	7.867551	12.99429	4	0	143	1496	26.90107	25.21898	18	0	152
m-rate (cancer)	Males	7769	31.23872	20.71804	28.77698	0	220.9945	1496	25.10603	8.304714	24.69617	0	64.18485
m-ratio (cancer)	Males	7769	114.9991	76.18444	105.5968	0	790.0212	1496	92.26116	30.1147	90.8595	0	231.511
Population	Females	7769	2869.272	4738.125	1291	111	46017	1496	11134.03	10045.72	7883.5	804	60349
Deaths (a.c.)	Females	7769	26.36041	41.19682	14	0	418	1496	75.9131	68.58967	52	3	397
m-rate (a.c.)	Females	7769	109.2031	46.83112	100.9398	0	552.3256	1496	70.3099	20.35932	68.0828	13.1579	212.3142
m-ratio (a.c.)	Females	7769	133.5369	55.74298	123.3041	0	609.2556	1496	86.0693	24.09801	83.23531	17.07777	234.0255
Deaths (cancer)	Females	7769	5.220749	9.270234	3	0	99	1496	17.90307	17.55498	12	0	120
m-rate (cancer)	Females	7769	18.93827	14.84767	17.24138	0	134.5291	1496	16.04026	7.096744	15.52677	0	65.32066
m-ratio (cancer)	Females	7769	109.6779	85.73326	100.2249	0	821.1254	1496	92.74087	40.5482	89.86648	0	411.9781

Notes: This table reports descriptive statistics on population, deaths, mortality rates, and mortality ratio at the municipality level, distinguishing between genders and whether the municipality under analysis belongs to the Land of Fires or the control group. Specifically, we report descriptive statistics on mortality for all causes of death and cancer-specific mortality in this table (*a.c.* stands for all causes, and *cancer* stands for cancer-type). *Deaths (a.c.)*, *m-rate (a.c.)*, and *m-ratio (a.c.)* stand for, respectively, the deaths, mortality rate, and mortality ratio for all causes of death. *Deaths (cancer)*, *m-rate (cancer)*, and *m-ratio (cancer)* instead stand for, respectively, the deaths, mortality rate, and mortality ratio for cancer.

cally, we report results on the residing population, deaths (in levels), and the two mortality indicators described above. The period considered is from 2003 to 2019.

We also re-run the same descriptive statistics just shown for a specific year and on a subset of the period from 2003 to 2019 (before and after the agreement's implementation for TdF), but the results do not change.[12]

For what concerns descriptive statistics on mortality for all causes of death, what emerges from Table 3.1 is that the number of deaths (number of "raw" deaths) in the municipalities that are part of the TdF is higher. By contrast, the mortality rate and the mortality ratio are higher in the rest of the region (true for both genders).

For the descriptives on cause-specific mortality, we considered only deaths whose primary cause of death was a malignant tumor. Specifically, we excluded from the analysis municipalities with less than two deaths for a specific sex since we did not observe the leading cause of death. Results align with those on mortality (all causes of death). No excess mortality is detected in TdF.

Failing to find any evidence of a correlation between affected municipalities and health outcomes – contrary to what was documented in the existing literature (see, for instance, ISS, 2020) – has two different explanations. First, unlike previous studies, our mortality indicators are not standardized to account for different age distributions. This is because, as explained at the beginning of this section, ISTAT does not release information by age for cause-specific deaths due to privacy concerns.

Second, given the substantial heterogeneity in terms of socio-economic characteristics featured by municipalities in Campania, some of the previous studies (see, for instance, Musmeci et al., 2015; Fusco et al., 2020) compared only municipalities that belonged to the same quartile of the deprivation index distribution.

We tried to tackle all these issues. We first tried to aggregate deaths for a specific municipality

[12]These descriptive statistics are not shown to save space but are available upon request.

across years and then compared our total number of deaths for cancer with those reported in Table 1 of page 6 of Attachment A of ISS (2020). Although we had a perfect match for cancer-specific deaths (in levels), our mortality ratio differed substantially from ISS (2020)'s, and this is probably because we cannot standardize our indicator by age.

Next, since we have data on age-specific mortality (all causes of death), we checked whether age-specific mortality indicators were higher in affected municipalities. Table 3.2 reports these descriptive statistics.

What emerges is that there is an excess of mortality in TdF (for both genders) for age groups $45-64$, $65-84$, and $85+$. In addition, the difference in mortality between TdF and the rest of the region increases as age groups increase. In contrast, there seems to be an excess of mortality in untreated municipalities when considering the mortality ratio for men $0-44$. Similar results are obtained when the analysis is restricted to only the provinces of Naples and Caserta (results not shown).

These results align perfectly with the results of ISS (2020), which document a deficit in cancer mortality in the age group $0-19$ in TdF compared to the rest of the region.

Despite these results appearing puzzling – given that studies document that higher exposure to hazardous waste increases the likelihood of congenital malformations (Geschwind et al., 1992; Dolk et al., 1998; Ravindra et al., 2021) – there are several reasons why we only find an effect for individuals aged 45 or older. Although we cannot rule out, at this stage of the analysis, the influence of healthcare service availability in TdF, one possible explanation for the descriptive results in Table 3.2 is that many types of cancer take years after (prolonged) exposure to manifest. Consequently, younger individuals may not yet show effects because they have been exposed to toxic pollutants for a shorter period.

Another concurrent explanation relates to data limitations. Unfortunately, due to data unavailability, we consider only one outcome: mortality. Although several of the aforementioned studies document an increase in cancer-related mortality, mortality itself is an extreme outcome.[13] Indeed, while an association between municipalities in TdF and increased cancer mortality is observed, other health effects of pollutant exposure cannot be ruled out. Thus, the absence of an observed effect on mortality for the age group $0-19$ does not necessarily imply that younger and more vulnerable individuals are immune to exposure to toxic waste. As reported in ISS (2020). However, there is no excess incidence or mortality of cancer among people aged $0-19$, this age group shows a higher incidence of leukemia and asthma episodes in TdF compared to the rest of the region.

Finally, a third potential explanation is that the results presented in Table 3.2 are for mortality from all causes. Despite an increase in overall mortality may be driven by higher cancer-related deaths, this aggregation could mask additional heterogeneity that we cannot capture. Unfortunately, as explained in the previous subsection, we do not have age-specific mortality data by cause of death.

To corroborate our findings, we performed some correlation tests between the age-specific mortality rate (all causes of death) and the cancer-specific mortality rate (results are not shown to save space). A negative correlation emerges between the cancer-specific mortality rate and the age-specific mortality rate (all causes) for the age groups $0-19$ and $20-44$. In contrast, a positive correlation for the age groups $45-64$, $65-84$, and $85+$ is detected. These results hold for both sexes.

[13] The incidence and mortality rates of malignant tumors vary significantly by age and gender. For example, among males aged $0-49$, the most frequent type of cancer is testicular cancer, while among those aged $50-69$, prostate and lung cancers are the most prevalent. Although testicular cancer does not rank among the five most fatal cancers, lung cancer has the highest mortality rate among males, particularly in the $50-69$ age group (ARITUM, 2015).

Table 3.2: Descriptive statistics – by age groups.

Age group: 0-19

		No Land of Fires						Land of Fires					
	Sex	**N**	**Mean**	**SD**	**Median**	**Min**	**Max**	**N**	**Mean**	**SD**	**Median**	**Min**	**Max**
Population	Males	7769	612.7704	1074.208	254	9	11837	1496	2848.538	2636.777	1999.5	152	16519
Deaths	Males	7769	0.2087785	0.5731013	0	0	7	1496	1.004011	1.42739	1	0	10
m-rate	Males	7769	3.535732	14.84865	0	0	384.6154	1496	3.443423	5.377215	1.249922	0	55.55556
m-ratio	Males	7769	102.2825	419.8913	0	0	10188.56	1496	98.74363	152.8745	35.60488	0	1763.803
Population	Females	7769	580.9762	1023.694	241	8	11533	1496	2699.753	2503.607	1880	125	15593
Deaths	Females	7769	0.1365684	0.4697465	0	0	7	1496	0.6624332	1.074648	0	0	8
m-rate	Females	7769	2.388185	13.81471	0	0	526.3158	1496	2.363374	4.547509	0	0	73.52941
m-ratio	Females	7769	100.6407	614.8609	0	0	25748.54	1496	99.88351	204.1194	0	0	4406.316

Age group: 20-44

		No Land of Fires						Land of Fires					
	Sex	**N**	**Mean**	**SD**	**Median**	**Min**	**Max**	**N**	**Mean**	**SD**	**Median**	**Min**	**Max**
Population	Males	7769	967.5223	1599.148	426	26	16286	1496	4020.967	3583.831	2888	240	21566
Deaths	Males	7769	0.8253314	1.719368	0	0	28	1496	3.675134	4.080635	2	0	27
m-rate	Males	7769	8.858617	17.622	0	0	322.5807	1496	8.773204	7.314707	8.08789	0	64.93507
m-ratio	Males	7769	101.2007	202.9663	0	0	4023.934	1496	100.0828	82.7979	95.40254	0	717.4938
Population	Females	7769	966.2645	1620.296	415	19	16701	1496	4124.002	3740.185	2886.5	238	22992
Deaths	Females	7769	0.3916849	0.9239283	0	0	12	1496	1.756016	2.143349	1	0	15
m-rate	Females	7769	4.082911	12.64052	0	0	357.1429	1496	4.122156	4.709822	3.395589	0	38.46154
m-ratio	Females	7769	98.92269	312.7622	0	0	10193.79	1496	99.59053	113.6615	81.91846	0	932.7458

Age group: 45-64

		No Land of Fires						Land of Fires					
	Sex	**N**	**Mean**	**SD**	**Median**	**Min**	**Max**	**N**	**Mean**	**SD**	**Median**	**Min**	**Max**
Population	Males	7769	712.9347	1151.551	322	30	10958	1496	2632.311	2398.116	1852	187	16041
Deaths	Males	7769	3.882997	6.872858	2	0	82	1496	15.84358	15.15969	11	0	93
m-rate	Males	7769	55.31166	48.70085	49.83389	0	441.1765	1496	59.50564	23.97447	58.12654	0	166.6667
m-ratio	Males	7769	96.74432	84.88906	87.09647	0	812.9427	1496	103.9268	40.44073	101.7674	0	275.1151
Population	Females	7769	739.4257	1244.205	329	29	12003	1496	2754.989	2548.668	1869	191	16910
Deaths	Females	7769	2.129489	4.118353	1	0	45	1496	9.28008	9.361904	6	0	59
m-rate	Females	7769	28.32308	35.75837	21.57497	0	555.5555	1496	33.19248	16.2847	32.62042	0	147.7832
m-ratio	Females	7769	91.58456	115.6417	70.12996	0	1854.847	1496	107.3323	52.40846	105.5431	0	453.7037

Age group: 65-84

		No Land of Fires						Land of Fires					
	Sex	**N**	**Mean**	**SD**	**Median**	**Min**	**Max**	**N**	**Mean**	**SD**	**Median**	**Min**	**Max**
Population	Males	7769	404.5572	611.9474	215	22	6736	1496	1124.9	1033.678	775	98	6358
Deaths	Males	7769	14.58193	23.14713	8	0	247	1496	44.95455	40.557	30	1	222
m-rate	Males	7769	359.922	154.4277	350	0	1388.889	1496	402.5764	96.39275	398.0184	65.78947	877.193
m-ratio	Males	7769	95.60423	40.77451	93.24802	0	424.7061	1496	106.7785	23.56188	106.6683	18.0847	226.0729
Population	Females	7769	499.0246	777.4198	261	33	7945	1496	1397.295	1276.538	953	146	7221
Deaths	Females	7769	11.60046	19.62626	6	0	205	1496	37.88369	34.12615	26	0	173
m-rate	Females	7769	224.6722	107.2783	220.5882	0	1153.846	1496	274.0214	71.64466	267.7952	0	631.5789
m-ratio	Females	7769	91.11503	43.53151	89.67188	0	450.6977	1496	110.8294	27.03294	109.3709	0	246.0932

Age group: 85+

		No Land of Fires						Land of Fires					
	Sex	**N**	**Mean**	**SD**	**Median**	**Min**	**Max**	**N**	**Mean**	**SD**	**Median**	**Min**	**Max**
Population	Males	7767	39.88837	52.99963	25	0	619	1496	67.34559	66.18577	45	2	482
Deaths	Males	7767	6.987254	9.850007	4	0	121	1496	13.4004	13.16784	9	0	81
m-rate	Males	7760	1776.747	1003.32	1692.308	0	10000	1496	2034.385	787.0043	2000	0	6666.667
m-ratio	Males	7760	97.05719	53.97181	93.84814	0	559.3235	1496	111.0978	41.94466	109.6793	0	349.0173
Population	Females	7769	83.5809	114.7536	51	2	1204	1496	157.988	152.7503	103	10	956
Deaths	Females	7769	12.1022	17.69113	7	0	189	1496	26.33088	25.15008	17	0	159
m-rate	Females	7769	1443.123	618.6934	1400	0	7692.308	1496	1709.957	513.5118	1666.667	0	5454.545
m-ratio	Females	7769	94.65116	39.51288	92.5532	0	428.7675	1496	112.0705	31.74622	110.6303	0	346.8987

Notes: This table reports descriptive statistics on population, deaths, mortality rates, and mortality ratio at the municipality level, distinguishing between genders, age groups, and whether the municipality under analysis belongs to the Land of Fires or the control group. *m-rate*, and *m-ratio* stand for, respectively, mortality rate and mortality ratio.

The documented correlation suggests that probably the excess mortality found for individuals aged $45+$ residing in affected territories can be driven by these individuals dying more from cancer.

Given the correlation found, we also performed some rank correlation tests (Kendall's τ and Spearman's ρ). Again, regardless of the test used or whether considering males or females, there is a lack of association between the cancer-specific mortality rate and mortality rate (all causes) for the $0-19$ age group (except in some cases, where a negative and statistically significant association is found). In contrast, a positive and statistically significant association between the mortality rate (all causes) for the other age groups and the cancer-specific mortality rate is detected.

Table 3.3: Descriptive statistics on sociodemographic characteristics.

		No Land of Fires						Land of Fires					
	Sex	**N**	**Mean**	**SD**	**Median**	**Min**	**Max**	**N**	**Mean**	**SD**	**Median**	**Min**	**Max**
% population 0-19	Males	7769	20.25	3.78	20.25	7.37	35.07	1496	26.45	2.95	26.45	19.43	34.78
% population 20-44	Males	7769	33.49	3.34	33.66	17.96	44.06	1496	36.95	2.31	37.00	28.00	43.75
% population 45-64	Males	7769	26.67	2.96	26.73	17.91	36.65	1496	24.79	2.50	25.01	18.05	31.38
% population 65-84	Males	7769	17.43	4.13	17.19	7.41	37.43	1496	11.10	2.55	10.99	4.64	20.38
% population 85+	Males	7767	2.16	1.33	1.85	0.00	10.09	1496	0.71	0.38	0.62	0.11	2.71
% population 0-19	Females	7769	18.49	3.55	18.53	5.41	34.18	1496	24.12	3.08	24.12	15.19	32.50
% population 20-44	Females	7769	31.44	3.77	31.64	11.91	43.32	1496	36.53	2.80	36.55	27.89	44.77
% population 45-64	Females	7769	25.44	2.99	25.60	15.21	36.04	1496	24.55	2.87	24.81	16.98	31.04
% population 65-84	Females	7769	20.49	4.65	20.09	9.08	40.23	1496	13.25	2.89	13.00	6.26	22.68
% population 85+	Females	7769	4.15	2.15	3.78	0.41	18.47	1496	1.55	0.70	1.41	0.28	4.76
Altitude zone	All	7769	2.75	1.23	3.00	1.00	5.00	1496	4.77	0.54	5.00	3.00	5.00
Altitude (in meters)	All	7769	375.44	233.35	370.00	2.00	1090.00	1496	59.38	50.51	42.50	3.00	340.00
Coastal	All	7769	0.12	0.32	0.00	0.00	1.00	1496	0.06	0.23	0.00	0.00	1.00
Coastal area	All	7769	0.23	0.42	0.00	0.00	1.00	1496	0.26	0.44	0.00	0.00	1.00
Level of urbanization	All	7769	2.60	0.59	3.00	1.00	3.00	1496	1.49	0.54	1.00	1.00	3.00
Per capita taxable income	All	7769	6992.65	3581.19	6777.18	51.39	172479.40	1496	6199.72	1457.71	6059.04	849.12	12791.34
Population density per km2	All	7769	462.32	1202.70	119.74	6.20	13067.40	1496	2371.37	2030.05	1731.72	83.25	9840.77

Notes: This table reports descriptive statistics on the municipality's population distribution by age and gender and on other socio-economic characteristics that we observe at the municipality level.

Lastly, Table 3.3 reports descriptive statistics on different socio-economic characteristics. Again, we considered the period from 2003 to 2019 and split results, when necessary, according to gender.

What emerges is that, as previous studies pointed out, there is substantial heterogeneity across municipalities in terms of socio-economic characteristics. Besides being larger in terms of population size, treated municipalities are, on average, closer to the sea, exhibit a higher level of urbanization (please note that *level of urbanization*$= 1$ means that it is a city), and feature a much higher population density, but are poorer. Thus, failing to account for such differences may bias the results.

Therefore, we performed a battery of descriptive regressions to compare municipalities with similar socio-economic characteristics rather than adjusting our mortality indicators for deprivation indexes.

Specifically, we regressed the cancer-specific mortality rate on a binary variable indicating if the municipality is situated in the Land of Fires. Further, we included as controls all the socio-economic characteristics reported in Table 3.3.

To avoid endogeneity issues related to the simultaneity of population-related variables on the right and left-hand sides (as explained in the previous subsection), we took the averages of these variables computed from 1980 to 1989. Further, since we do not observe age-specific mortality from cancer, we also included the averages computed from 1980 to 1989 of the distribution of deaths (all causes) by age.

Standard errors were clustered at the municipality level to account for the potential serial correla-

tion of the error term.

Results are reported in Appendix A. Specifically, Table A.1 reports the results for the mortality rate from cancer for males; whereas Table A.2 shows the results for females. Keeping these controls fixed, we find that the cancer mortality rate was higher in affected municipalities. This result holds for both genders.

We also re-run the same analysis reported in Tables A.1 and A.2 using only the provinces of Naples and Caserta (rather than the entire region). Results are the same. When controlling for covariates, living in TdF is associated with a higher mortality rate from cancer. Again, this result applies to males and females. To save space, these results are omitted but are available upon request.

Overall, these findings suggest that there is an excess of cancer-specific mortality in TdF compared to the rest of the region. The fact that the mortality rates reported in Table 3.1 are higher outside than inside TdF is probably because we are not standardizing the mortality indicators. Thus, we fail to compare municipalities that feature similar age-specific mortality rates. For this reason, we believe it is more appropriate to exploit causal inference methods that compare treated and untreated units and directly account for differences in observable characteristics to estimate the causal effect.

3.5 Empirical Strategy

3.5.1 Matching Estimators

To estimate the causal effect of the Land of Fires, we decided to consider only the mortality rate from cancer per 10,000 inhabitants.[14]

In addition, we chose to consider only cancer-specific mortality rather than mortality from all causes of death because there is sufficient medical evidence linking cancer mortality to hazardous waste. While considering all-cause mortality in the descriptive analysis allowed us to account for potential differences in mortality due to age distribution, our empirical strategy now enables us to directly control for age-group differences, even when focusing on cancer-specific mortality.

Although estimating the municipal-level effect of living in affected territories may be interesting, the lack of a randomized experiment makes this task impossible Holland (1986).

For this reason, we consider two quantities commonly used in the program evaluation literature to estimate the average effect of the policy: the *Average Treatment Effect* and the *Average Treatment Effect on the Treated*.

Using the potential outcomes framework first introduced by Neyman (1923) and later formalized by Rubin (1974), we define the ATE as follows:

$$ATE = \mathbb{E}\left[y_i(1) - y_i(0)\right] \tag{3.1}$$

Let D_i be a binary treatment indicator that takes the value 1 if municipality i belongs to the Land of Fires and 0 otherwise, then $y_i(1)$ represents the potential outcome for unit i had it received the treatment (i.e., $D_i = 1$), and $y_i(0)$ the potential outcome had it not received the treatment (that is, $D_i = 0$).

[14] We chose not to estimate the impact on the mortality ratio, as it contains almost the same information as the mortality rate. Furthermore, for ease of interpretation, it is clearer to understand the implications of a positive (or negative) and statistically significant effect on the mortality rate, while the same does not apply to the mortality ratio due to its nature.

The ATE measures the expected average effect of the policy on the entire population of interest. In this context, it represents the expected difference in (cancer-specific) mortality rates if the Campania as a whole region had been exposed to uncontrolled dumping and burning of waste compared to a scenario in which no municipality had been exposed to this phenomenon.

On the other hand, the ATT estimates the expected average effect of the policy under analysis on the treated subpopulation:

$$ATT = \mathbb{E}\left[y_i(1) - y_i(0)|D_i = 1\right] \tag{3.2}$$

Different estimators for the parameters in equations (3.1) and (3.2) exist, each relying on different identifying assumptions.

Given the context under analysis, we decided to rely on one of the most commonly used estimators: the *inverse-probability-weighted regression adjustment*, also known as *doubly-robust* estimator. We also re-ran the analysis using the *propensity-score matching* as a robustness check.

Both methods allow us to account for the fact that treatment assignment is not purely random, but there is selection based on observable characteristics.

Specifically, propensity-score matching estimates the missing potential outcome for each unit by averaging the outcomes of similar subjects who received the opposite treatment. The similarity between treated and untreated units is assessed using an estimate of the propensity score. The average treatment effect is then calculated by averaging the differences between each subject's observed and counterfactual outcomes (Caliendo and Kopeinig, 2008; Abadie and Imbens, 2009).

By contrast, doubly robust (DR) estimators apply probability weights – obtained via estimation of the inverse of the propensity score – to derive outcome-regression parameters that adjust for the selection problem. These adjusted parameters are then used to compute averages of predicted outcomes for treated and control units. The differences between these averages represent the estimate of the ATE (Robins and Rotnitzky, 1995; Robins et al., 1995; Wooldridge, 2007).

Compared to propensity-score matching, doubly robust estimators "usually enjoy additional robustness against model-misspecifications" (Callaway and Sant'Anna, 2021, p. 207). This additional robustness comes from the fact that the estimator is consistent as long as one of the two models (the inverse probability weighting or the outcome regression/regression adjustment) is correctly specified.[15]

The propensity-score matching and the IPWRA method rely on two key identifying assumptions. The first is the *strong ignorability* or *unconfoundess* assumption, which requires potential outcomes to be independent of treatment assignment after conditioning on observable characteristics. Formally,

$$(y(1), y(0) \perp\!\!\!\perp D|X)$$

Where X is a vector of pre-treatment observable characteristics.

Simply put, the strong ignorability assumption states that conditioning on observable characteristics is sufficient for treatment assignment to be as good as random (Caliendo and Kopeinig, 2008; Wooldridge, 2010; Abadie and Cattaneo, 2018).

The second assumption is known as *overlap assumption*, which requires the propensity score (i.e., the conditional probability of receiving the treatment), $\mathbb{P}(D = 1|X)$, to be bounded away from 0 and

[15] For a review of how the doubly robust estimator of the ATE works, interested readers can find more details in Wooldridge (2007, 2010).

1.[16]

To give the intuition, the overlap assumption ensures that municipalities with the same values of observable characteristics have a positive probability of being both treated and untreated.

Although various methods exist to assess whether the overlap assumption holds in the given context, the unconfoundedness assumption is inherently untestable. We rely on the institutional setting to justify why it is likely to hold in our case.

As explained in Section 3.2, assignment to the treatment group is an endogenous process, as Patto per la Terra dei Fuochi is an agreement between municipalities, regional authorities, and central authorities.

However, at least concerning the outcomes considered in this study (i.e., mortality rates), the treatment assignment mechanism is likely to be exogenous once observable characteristics are controlled for. Indeed, while factors such as the number of open-air (illegal) landfills, toxic waste combustion, air pollution from burning waste, attitudes toward littering and illegal disposal, water contamination, and the mapping and requisitioning of contaminated agricultural lands were undoubtedly among the criteria used by policymakers to classify a municipality as part of Terra dei Fuochi, neither Patto per la Terra dei Fuochi nor DL 136/2013 "Terra dei Fuochi" explicitly referenced cancer incidence or mortality rates as a criterion for treatment assignment.

In other words, suppose that two municipalities, A and B, had different mortality rates before 2013, with municipality A experiencing a significantly higher mortality rate than B. This difference would not determine whether municipality A would be classified as part of Terra dei Fuochi starting in 2014.

Furthermore, to mitigate bias arising from omitted observable characteristics that may affect both mortality rates and treatment status, we controlled for various socioeconomic characteristics. Specifically, these include all the variables reported in Table 3.3. Additionally, we controlled for total number of deaths (all causes), the distribution of deaths by age, and year fixed effects. Controlling for year fixed effects might capture macroeconomic shocks that may affect all municipalities belonging to the same treatment group similarly.

However, because control variables should not be affected by treatment, we used pre-treatment values for time-varying characteristics (i.e., population and deaths in levels, along with their age distributions). Specifically, we used averages computed from 1980-1989.[17]

Regarding statistical inference, we used the Abadie-Imbens robust standard errors when estimating the ATE via propensity-score matching (Abadie and Imbens, 2006).[18] Conversely, when considering the IPWRA, we used cluster robust standard errors at the municipality level to account for the potential serial correlation in the mortality rate.

3.5.2 Threats to Identification

The program evaluation literature typically assumes that an independent and identically distributed random sample is drawn from the population. Random sampling implies the *stable unit treatment value assumption (SUTVA)*, which is sufficient to define the parameters in equations (3.1) and (3.2). For SUTVA

[16] A weaker version of the unconfoundess and overlap assumptions are sufficient for identification of the ATT (Caliendo and Kopeinig, 2008).

[17] This choice is motivated by evidence suggesting that criminal activities related to illegal waste disposal began in the late 1980s.

[18] Unfortunately, for matching estimators, a cluster-robust estimator of the variance-covariance matrix does not exist. Further, standard bootstrap procedures usually fail for these estimators (Abadie and Imbens, 2008).

to be valid, two conditions must be met: i) outcome of unit i is not affected by the treatment status of unit $j \neq i$ (also known as *no interference*); ii) treatment consistency.

One possible violation of consistency occurs when there are different levels of intensity of treatment, leading to different potential outcomes. The consistency of treatment is violated if these different potential outcomes are not explicitly accounted for in the analysis.

In the context under analysis, a violation of the consistency condition can occur when different sources and levels of pollutants affect outcomes differently, leading to different treatment intensities/levels. For example, if individuals living near different landfills are exposed to different types and quantities of pollutants compared to those living farther away, but this source of treatment heterogeneity is not explicitly incorporated into the analysis (i.e., failing to account for differential effects due to landfill proximity). The definition of treatment becomes ambiguous and consistency is violated.

Although a violation of treatment consistency is likely in our context – particularly when considering whether the effect varies by proximity to illegal dumping sites – this analysis focuses solely on understanding whether individuals residing in the Land of Fires experience higher cancer mortality rates than those living outside, regardless of their distance from illegal dumping sites. Further, despite investigating potential heterogeneity in mortality based on landfill proximity, which would be valuable, such an analysis is not feasible due to data limitations; only aggregate data are available due to privacy concerns.

Regarding the assumption of no interference, in this context, it implies that the potential outcome (i.e., mortality rates) of individuals living in the municipality i is not affected by whether the municipality $j \neq i$ is located in TdF. This assumption is likely to hold, as municipalities operate relatively independently regarding healthcare service provision and environmental policy enforcement. However, there may be exceptions.

Two potential violations of the SUTVA may arise due to (negative) spillover effects. These spillovers result from waste combustion, as the resulting toxic cloud can spread to neighboring municipalities. To illustrate how SUTVA can be violated in this context, we provide two examples and adopt the terminology used in the causal inference literature to model spillovers (e.g., see Vazquez-Bare, 2023).

The first scenario where SUTVA may be violated occurs when spillovers exist from treated to untreated units. Consider *Castel Volturno* and *Cancello ed Arnone*, two neighboring municipalities in the province of Caserta. While Castel Volturno is officially recognized as part of the Land of Fires, Cancello ed Arnone is not. A violation of SUTVA can occur if large toxic clouds generated by waste combustion in Castel Volturno spread through the air to Cancello ed Arnone. If residents of Cancello ed Arnone – particularly those near the border – are exposed to carcinogenic substances leading to an increased incidence of malignant tumors and higher mortality rates, then interference exists. If such a violation is not adequately accounted for, the estimators for the ATE and ATT can be proven to be biased.

The second scenario is analogous but involves spillovers among treated units. Consider *Acerra* and *Caivano*, two municipalities within TdF. A potential violation of SUTVA occurs if toxic clouds from waste combustion in Acerra propagate to Caivano, thereby affecting its residents' incidence and mortality rates from cancer. Specifically, the mortality rate in Caivano would be impacted twice: first, by local open-air (both legal and illegal) landfills located on its territories and second, by air pollution spilling over from neighboring treated municipalities.

Disentangling these two effects in the second scenario is challenging. The difficulty arises from the inability to systematically record the amount and types of toxic waste at each site. Even with precise

geolocation data on open-air landfills, there would still be no reliable information on active fires. We acknowledge that this may be a limitation of the current study, as the existing causal inference literature on spillover effects explicitly requires modeling which units are subject to spillovers and which are not, to have an unbiased estimator of the average causal effect of the policy.

In contrast, the first scenario is easier to address. One possible solution is to exclude municipalities not part of the Land of Fires but are immediate neighbors from the control group. Alternatively, the analysis could be re-run by treating these latter municipalities as if they were treated.

To assess whether a violation of the no-interference assumption is likely in our context, we will re-run the main analysis, excluding municipalities from the control group that are not part of the Land of Fires but are immediate neighbors. If no spillover effects are present, we expect the estimated ATE and ATT to be similar to those obtained using the remaining 460 untreated municipalities. However, if the estimated ATE and ATT change substantially, mainly if they are larger than the main analysis, where these neighboring municipalities were still part of the control group, this could have two implications. First, it would suggest that the Land of Fires extends beyond the 90 municipalities officially recognized by the central government. Second, it would indicate that these municipalities, affected by negative spillover effects, are unsuitable controls and, therefore, should be excluded from the analysis.

While the SUTVA is sufficient to define the estimands of interest, as explained in the previous subsection, two additional assumptions are required to estimate the parameters in equations (3.1) and (3.2) via either IPWRA or propensity-score matching.

Regarding the overlap and common support assumption, as Heckman et al. (1997) and Caliendo and Kopeinig (2008) point out, a violation of this assumption is a significant source of bias in analyses based on matching estimators. To avoid comparing "the incomparable", "an important step is to check the overlap and the region of common support between treatment and comparison groups" (Caliendo and Kopeinig, 2008, p.45). Most common statistical software directly checks whether the overlap assumption is satisfied and excludes from the estimation units for which this assumption is violated. However, when estimating the ATE and the ATT, we follow Caliendo and Kopeinig (2008) and inspect the densities of the propensity score in the treated and untreated groups.

Although, as explained in the previous subsection, the treatment is likely exogenous with respect to cancer mortality rates, matching methods have an important limitation. Both IPWRA and propensity-score matching require that, once having conditioned on pre-treatment observable characteristics, other unobservable variables that affect the outcome of interest are unrelated to treatment assignment. As DiPrete and Gangl (2004) point out, treatment assignment will not generally be independent of unobserved variables in non-experimental designs.

Even though we control for different socioeconomic characteristics, making the validity of the unconfoundedness assumption more likely, data limitations prevent us from ruling out the possibility that any positive and statistically significant effect detected in municipalities situated in the Land of Fires is not attributable to the different availability of healthcare services in these municipalities. Furthermore, with the current data, we cannot control for different substances disposed of on the grounds or for other potential sources of pollution (e.g., the presence of high-carbon-emission industries).

Failure to account for these unobserved variables could compromise the validity of our analysis. To validate our findings, we will perform two robustness checks. First, we will re-run the entire analysis using only the provinces of Naples and Caserta. Excluding municipalities that are geographically farther away and comparing treated municipalities with untreated ones from the same province should

improve comparability by reducing potential differences in both observable and unobservable characteristics.

The second robustness check will involve exploiting the *Rosenbaum bounds* (Rosenbaum, 2002). These bounds are used in sensitivity analysis to assess how hidden bias due to unobserved confounders might affect causal inference in observational studies, particularly in matching methods.

Specifically, Rosenbaum bounds allow us to assess how sensitive the ATT estimates are to unobserved confounding. If an unobserved confounding variable, W, does not affect treatment assignment, then the bounds equal the significance level estimated using the matching method. However, as the impact of W on treatment assignment strengthens, the confidence intervals for the test of the absence of an effect of the treatment on the outcome widen. For a more detailed discussion, see Rosenbaum (2002) and DiPrete and Gangl (2004).

Suppose endogeneity due to omitted unobserved confounders is indeed a problem. In that case, we should expect the estimated effect to differ drastically when re-running the analysis using only the provinces of Caserta and Naples. Similarly, when considering Rosenbaum bounds, we should observe that our analysis is particularly sensitive when increasing by a small amount the potential impact of W on D.

3.6 Empirical Results

3.6.1 Main Findings

We estimate the parameters in (3.1) and (3.2) using the inverse probability-weighted regression adjustment estimator. Compared to other methods, doubly robust estimators enjoy additional robustness against model misspecification. The reason is that the requirement for consistency of the IPWRA estimator is only one of the two models being correctly specified (i.e., IPW or RA).

Table 3.4 reports the estimates of the ATE and ATT obtained via IPWRA estimand when using as a dependent variable the cancer-specific mortality rate for $10,000$ inhabitants. The propensity score model is estimated using a logit, and we opt for a linear outcome model. Cluster-robust standard errors, clustered at the municipality level, are reported to account for the potential serial correlation in the mortality rate.

Results are presented separately for males and females. The choice to report results separately is dictated by the existence of medical studies documenting a differential incidence and mortality of specific malignant tumors by gender (e.g, see ARITUM, 2015; Selvaraj et al., 2025). Since we cannot account for gender-based differences in cancer incidence, nor do we distinguish between cancer types, reporting results without differentiating by sex may obscure this heterogeneity.

As explained in the previous section, many covariates are included to make the unconfoundness assumption likely valid. If, on one side, this assumption is not directly testable, conditioning on more relevant covariates should help make the assumption of independence between potential outcomes and treatment more plausible, i.e., strengthen the plausibility of the unconfoundedness assumption. In particular, specification (1) includes as controls population and mortality (all causes of death) in levels and their respective distribution by age. To avoid possible simultaneity issues or the fact that these variables have been affected by exposure to environmental contamination, we took the averages of these variables computed using the years 1980 to 1989.

Table 3.4: IPWRA results for cancer-specific mortality rate

	Males			Females		
	(1)	(2)	(3)	(1)	(2)	(3)
ATE (TdF vs non TdF)	4.55***	1.88**	1.53**	6.90***	7.33***	6.38***
	(0.84)	(0.84)	(0.74)	(0.67)	(0.64)	(0.62)
ATT	2.28***	4.50***	3.25***	1.80***	3.54***	3.08***
	(0.85)	(0.94)	(0.75)	(0.46)	(0.74)	(0.62)
Pop. & Death (in levels)	Yes	Yes	Yes	Yes	Yes	Yes
Pop. & Death (distr. by age)	Yes	Yes	Yes	Yes	Yes	Yes
X	No	Yes	Yes	No	Yes	Yes
Year FE	No	No	Yes	No	No	Yes
Observations	7,837	7,743	7,806	7,956	7,515	7,590

Notes: Coefficients are reported with cluster-robust robust standard errors in parentheses clustered at the municipality level to account for the potential serial correlation in the mortality rate. The outcome of interest is the cancer-specific mortality rate. Estimates are performed via IPWRA. *Pop.* stands for population and *distr* for distribution. **X** is a vector including altitude zone, altitude (in meters), a dummy indicating whether it is a coastal town, coastal area, level of urbanization, income per capita, and population density. ***, **, and * denote significance at 1%, 5%, and 10%, respectively.

In column (2), we report results obtained by conditioning also on the other socioeconomic characteristics reported in Table 3.3 (i.e., income, population density, municipality's classification, etc.). Lastly, column (3) shows the results obtained when also controlling for year fixed effects.

Before commenting on the estimates of the ATE and ATT reported in Table 3.4, it is essential to explain why the sample size varies between specifications. The variation arises because we had to exclude more than $1,000$ observations from the estimation due to a violation of the treatment overlap assumption, which is necessary for estimating the propensity-score weights.

As explained in the previous section, modern statistical software directly excludes observations for which there is no common support and, therefore, where the overlap assumption is violated. However, following the suggestions in Caliendo and Kopeinig (2008), we inspect the densities of the (estimated) propensity score in both groups and find no differences in the region of common support.[19] The discarded observations are those with a very small propensity score value (≈ 0). Additionally, municipalities excluded due to a violation of the overlap assumption belong to the provinces of Avellino and Salerno – municipalities located farther from TdF and more likely to differ in both observable and unobservable characteristics. For the sake of space, we do not report these results here, but they are available upon request.

Regarding the estimates of the ATE, Table 3.4 shows that the impact is positive and statistically significant for both sexes. People residing in affected areas have higher mortality rates than those living in the rest of the region, assuming the entire Campania region had been exposed to (toxic) waste disposal. These results hold across all specifications considered.

[19] A potential issue that could threaten our results would arise if, for instance, treated units had observations in the region $[0.1, 0.9]$. Still, there were no observations in the region $[0.6, 0.8]$ for untreated units. This is not the case in our analysis.

The same applies to the ATT estimates. Had municipalities in TdF not been affected by hazardous waste and fires, cancer mortality would have been lower. Again, these results hold across specifications and for both sexes.

In particular, when considering the ATE, we observe that for males, the magnitude decreases substantially from specification (1) to (2) but remains relatively stable when controlling for year fixed effects. The reduction in magnitude between columns (1) and (2) is likely due to omitted variable bias, where failing to control for variables correlated with both treatment status and covariates included in the specification (1) led to an overestimation. By contrast, the point estimates for females remain stable across all three specifications.

While a larger differential positive effect for females would be consistent with the findings of ISS (2020) – potentially explained by differences in cancer types, incidence rates, and survival rates between sexes (ARITUM, 2015) – the ATT estimates are quite similar for males and females. This holds across all specifications considered.

In general, across all specifications and for both males and females, residents of TdF experience adverse health effects compared to both the rest of the region and a counterfactual scenario in which none of these municipalities had been affected by the waste phenomenon.

These results align with previous studies documenting a positive association between cancer mortality and living in TdF (e.g., see ISS, 2020; Cafieri and Feoli, 2023; Fazzo et al., 2023). In addition to the higher cancer mortality rate in TdF, previous studies also report a higher incidence of malignant tumors in these areas compared to the rest of the region (ISS, 2020).

Although we lack detailed information on the type and quantity of toxic waste disposed of in these areas – limiting our ability to investigate potential explanatory channels – prior studies have documented the presence of dioxins and heavy metals in these regions (e.g., see ARPAC, 2017). Moreover, Fazzo et al. (2023) show that cancer-specific mortality is higher in municipalities with a high waste risk index.

Therefore, the higher incidence and mortality rates from cancer documented in TdF are consistent with international epidemiological and medical studies showing that exposure to waste combustion, hazardous waste disposal, and living near landfills negatively affect health outcomes (Jarup et al., 2002; Garcia-Perez et al., 2013; WHO, 2024).

3.6.2 Robustness Checks

We conducted a battery of sensitivity exercises and robustness checks to corroborate our findings.

First, we performed a placebo test. We selected a control group of municipalities geographically close to the treated group but without significant exposure to the Terra dei Fuochi phenomenon. Testing for policy effects in areas where no actual treatment occurred helps confirm the specificity and validity of the findings in Table 3.4.

To this end, we defined a control group consisting of all municipalities in the provinces of Naples and Caserta ($n = 72$) that were neither treated nor located on the borders of the Land of Fires (e.g., we excluded municipalities such as *Grazzanise* and *Torre Annunziata*).[20] We then set the probability of being treated to $Pr = \frac{1}{t}$, where $t \in \{3, 4, 5, 6\}$.[21] By varying t, we assess whether the results change

[20] We excluded border municipalities to avoid potential negative spillover effects from treated areas, as discussed in the previous section.

[21] The true fraction of treated municipalities is 16.15%. This is why we chose to let t to vary in $\{3, 4, 5, 6\}$

when increasing (decreasing) the percentage of untreated (treated) municipalities. Since the assignment to this new treatment is, by construction, unrelated to both actual treatment status and outcomes, if the effects observed in Table 3.4 are truly due to toxic waste exposure, we should find no significant effects in this placebo region.

The results of this placebo analysis, based on 1,000 permutations, are presented in Table A.3 in Appendix A. This table reports the estimated ATE and ATT for our preferred specification (column (3) in Table 3.4) obtained using IPWRA. Specifically, columns (1) to (4) progressively increase the number of untreated units, ranging from $\frac{2}{3}$ in column (1) to $\frac{5}{6}$ in column (4). With one exception (results for males when $p = 1/5$), all estimated ATE and ATT values are statistically indistinguishable from zero. Thus, confirming the specificity and validity of our findings.

As an ulterior check, we re-run the analysis reported in Table 3.4, restricting the sample to the provinces of Naples and Caserta instead of the entire region. Table A.4 in Appendix A replicates the specifications from Table 3.4, with the only difference being that the ATE and ATT estimates are now based solely on these two provinces.

The results in Table A.4 remain consistent with those in Table 3.4. However, while the ATE estimates are now smaller in magnitude, the ATT estimates are larger. The reduction in the ATE could be due to some of the untreated municipalities in these provinces also being affected by toxic waste issues, despite not being officially classified as part of the Land of Fires. This is particularly relevant if treated municipalities were previously matched with untreated municipalities farther from the affected areas when estimating the ATE for the entire region.

Conversely, the increase in ATT estimates may be because, by excluding more distant municipalities, we are now able to control for unobserved factors that influence both treatment and outcomes – factors that are more likely to be similar among municipalities within the same province. Additionally, the difference in estimates could be attributed to the ATT estimator relying on weaker assumptions for consistency.

As discussed in Section 3.5.2, a potential threat to our identification strategy is the presence of negative spillover effects generated from waste combustion, as toxic clouds may spread to neighboring municipalities. If such spillovers occur, the no-interference assumption would be violated.

Although we cannot entirely rule out spillovers among treated municipalities – acknowledging this as a limitation of our study– we can assess whether spillovers extend from treated to untreated municipalities. To do so, we re-run the specifications presented in Table 3.4, this time excluding untreated municipalities that are located at the border with the Land of Fires.

The results, presented in Table A.5, align with those reported in Table 3.4, suggesting that if spillover effects exist, they are likely to be a minor issue and do not pose a significant threat to our identification strategy.

To assess whether the introduction of the Terra dei Fuochi Decree enhanced the central government's effectiveness in preventing unlawful activities, thus mitigating adverse health effects, we re-run the analysis separately for the years before and after the policy's implementation. The results are presented in Table A.6 in Appendix A. Panel (a) reports estimates for 2003-2013, while Panel (b) presents results for 2014-2019. These findings closely mirror those of Table 3.4.

Regardless of the specification or time period considered, the estimated effects remain positive and statistically significant. The only exception is the estimate of the ATT for males, which is now larger in magnitude than those reported in Table 3.4 during the pre-treatment period. Still, it becomes smaller

and no longer statistically different from zero in the post-treatment period. However, this latter result may be driven by the shorter time frame considered rather than by the actual effect of the policy on male cancer mortality.

Two factors may explain why the ATE and ATT estimates remain statistically different from zero after the decree's introduction. First, the post-treatment period may be too short to detect significant effects, as policy impacts on health outcomes often take years to manifest. Second, despite efforts by local authorities, illegal waste dumping and burning continue, highlighting the need for more stringent interventions from the central government.

Next, we re-run the analysis reported in Table 3.4, but now using propensity-score matching as an estimator for the parameters in (3.1) and (3.2). The results are presented in Table A.7. Specifically, columns (1) to (3) show estimates obtained using the nearest-neighbor matching estimator, based on a one-to-one match (with replacement), to compute the counterfactual outcome for each municipality. The propensity-score estimates are derived using a logit model.

Before interpreting the results, it is important to highlight that the estimates for both males and females in column (1) are not directly comparable to those in the other two specifications. Although we always required matching a municipality with another municipality in the opposite treatment group, the software required 17 matches to compute the ATE and ATT in column (1). In contrast, a one-to-one match was used in all the other columns.

Disregarding the results in column (1), we observe that the estimates for females in Table A.7 closely align with those reported in Table 3.4. However, for males, the estimated ATT differs substantially from the ATE. While the estimated ATE is negative and statistically significant, the ATT is positive and statistically significant. Since propensity-score matching provides a consistent estimator for the ATT but not for the ATE, and because the ATT estimator requires weaker identifying assumptions for consistency, this discrepancy may be attributed to these factors.

Since the nearest-neighbor matching estimator based on a one-to-one match (with replacement) is only one possible matching algorithm, we tested the robustness of our findings by applying different Propensity Score Matching algorithms, following the recommendations of Caliendo and Kopeinig (2008). The performance of different matching algorithms varies case by case, largely depending on the data structure.

To this end, we first examine the impact of increasing the number of municipalities in the opposite treatment group to which a given municipality must be matched, rather than requiring a strict one-to-one match. The results are presented in Table A.8. For ease of comparability, column (1) reproduces column (3) from Table A.7, where a one-to-one match was used. In columns (2) to (4), we progressively increase the number of required matches from 2 to 4. The results indicate that our findings remain robust to changes in the number of matches.

Next, we assess the sensitivity of the specification in column (3) to variations in the caliper. The caliper defines the maximum allowable distance, in terms of the propensity score, between two observations for them to be considered potential matches. In Table A.7, all observations were potential matches regardless of their level of dissimilarity. The results of this analysis are shown in Table A.9. Specifically, we use a caliper of 0.1 in column (1), 0.25 in column (2), 0.5 in column (3), and 0.75 in column (4). The estimated parameters remain identical to those reported in column (3) of Table A.7. This consistency is because more than 1,000 observations had already been excluded due to violations of the overlap assumption, meaning that highly dissimilar municipalities had already been removed

from the estimation of the ATE and ATT.

Lastly, we examine the impact of using different propensity-score matching algorithms. Since propensity-score matching provides a consistent estimator for the ATT (but not for the ATE), we focus solely on the ATT for brevity. The results are presented in Table A.10.

Precisely, we assess whether the estimates reported in column (3) of Table A.7 (now shown in column (1) of Table A.10) are sensitive to the following modifications: Inclusion of observations that violated the overlap assumption (column (2)); One-to-one nearest-neighbor matching without replacement (column(3)); Kernel matching (column (4)); Local-linear regression matching (column (5)); Radius matching (column (6)); Imposition of common support by using common (column (7)) and trimming procedures (column (8)) when using one-to-one nearest neighbor matching estimators. In particular, results in columns (7) and (8) were obtained using one-to-one nearest-neighbor matching estimators. When trimming, we dropped the first 1% of the treated municipalities for which the estimated propensity score of the control municipalities is the lowest.

The results remain robust regardless of the matching algorithm used or the method of imposing common support. However, the magnitude of the estimates varies across methods, which is expected given that these algorithms construct the counterfactual outcome, $y(0)$, in different ways.

The only exceptions occur when estimating our model using one-to-one nearest-neighbor matching without replacement (for both sexes), local-linear regression (for males), radius matching (for both sexes), and when imposing common support (for males).

Unsurprisingly, one-to-one nearest-neighbor matching without replacement performs worse than its counterpart with replacement – since allowing replacement increases the likelihood of matching comparable units and reduces bias – estimates obtained via matching without replacement are also sensitive to the order in which observations are matched (Caliendo and Kopeinig, 2008). However, the fact that radius matching produces significantly different results is puzzling, as this method is generally designed to minimize the risk of poor matches.

Lastly, as discussed in Section 3.5.2, while we account for various socioeconomic factors, data unavailability prevents us from determining whether the higher mortality rate is driven by differences in the availability of healthcare services in the case of TdF or other unobserved confounders affecting both treatment assignment and cancer mortality. If such confounding factors exist, both Propensity-score Matching and IPWRA estimators may yield inconsistent estimates.

To evaluate the potential bias introduced by unobserved factors in our estimation of TdF's effect on cancer mortality, we perform a sensitivity analysis using Rosenbaum bounds (Rosenbaum, 2002).

Results are presented in Table A.11, and we report the $p-$values from the Wilcoxon signed-rank test for the ATT, allowing the level of hidden bias, Γ, to vary. This parameter represents the odds ratio of differential treatment assignment due to an unobserved confounder. Precisely, Table A.11 displays $p-$values under both positive and negative selection on unobservables (denoted as sig+ and sig-, respectively). Additionally, Hodges-Lehmann point estimates and 95% confidence intervals for the ATT are also reported.

For each Γ, it is computed a hypothetical significance level, named $p-$critical, which represents the bound on the statistical significance of the ATT in the presence of selection on unobservables (Rosenbaum, 2002; DiPrete and Gangl, 2004). When $\Gamma = 1$, this indicates no hidden bias (i.e., treatment assignment is entirely random given the observed covariates). By examining how the Rosenbaum bounds change when increasing Γ, the researcher can evaluate the extent to which unobservable confounders

would need to influence the results for the estimated ATT to be entirely attributable to nonrandom assignment.

Under negative selection, our results remain highly robust even for values of Γ that increase the log-odds of differential treatment assignment by 300%. Under positive selection, our findings also hold, with the critical level of Γ at which we would need to reconsider the conclusion of a positive TdF effect ranging between 2.25 and 2.50 for males and between 2.75 and 3 for females.

Overall, we can thus claim that the results shown in Table 3.4 are pretty robust to different sensitivity analyses and robustness checks.

3.6.3 Discussion

As expected, the analysis conducted in Subsections 3.6.1 and 3.6.2 demonstrates that failing to account for heterogeneity in socio-demographic characteristics across municipalities can bias the results. Given the substantial differences in observable characteristics across municipalities (Istat, 2020; Svimez, 2022), controlling for these variables is paramount as it helps compare similar municipalities in TdF with those in the rest of the region.

The absence of any adverse effects in Table 3.1, while reassuring on the one hand, on the other hand, could stem from the fact that we did not control for the age distribution, unlike previous studies. In fact, when estimating the ATE on cancer-specific mortality, a positive and statistically significant effect was detected for individuals residing in TdF. These findings hold for both sexes and across different robustness exercises, which also take into account the possibility of spillover effects and the existence of unobservable factors that may correlate both with treatment assignment and outcomes.

The adverse effect documented in previous subsections aligns perfectly with the existing descriptive evidence reported in the epidemiological literature on the effects of TdF (Trinca et al., 2001; Altavista et al., 2004; Fusco et al., 2020; ISS, 2020; Cafieri and Feoli, 2023; Fazzo et al., 2023). The increased incidence and mortality for specific malignancies can be explained by significant contamination of animals and humans living in affected areas (Maselli et al., 2010; Rivezzi et al., 2013; ISS, 2020; Fazzo et al., 2023).

In addition, our results are consistent with the international literature showing that exposure to waste combustion, hazardous waste disposal, and living near landfills have adverse health effects (Jarup et al., 2002; Garcia-Perez et al., 2013; WHO, 2024).

Together, these findings provide strong evidence for the detrimental health effects of illegal waste disposal, thus contributing to the ongoing debate of whether residing in the Land of Fires negatively affects health outcomes. In addition, they also provide further evidence to the international literature on the possible adverse effects of environmental contamination.

Despite the actions undertaken by the central and local authorities, the illegal burning and dumping of toxic waste continue.

These findings are particularly concerning when considering that, between 2017 and 2022 alone, the Campania region is estimated to have spent 30 million Euros annually (amounting to a total of 173 million Euros) on oncology services for its inhabitants through passive mobility only.[22]

Overall, the persisting environmental crisis and the increased incidence of specific malignancies

[22]Please, refer to https://ancicampania.it/aumentano-i-tumori-in-italia-in-campania-si-spendono-30-milioni-allanno/.

in these territories call for more drastic policy interventions, such as land decontamination, cancer screening programs, and more funds from the central government to local authorities to prevent and defeat these crimes.

3.7 Concluding Remarks

Uncontrolled and illegal hazardous waste disposal seriously threatens individuals' health. For this reason, from the beginning of the 2000s, numerous studies have tried to understand whether there were systematic differences in specific pathologies between individuals living in the Land of Fires and the rest of the region.

The present work's primary goal is to provide further evidence to the existing literature and to estimate the TdF's causal effect for the first time. To this end, we employ data from different administrative sources and consider cancer-specific mortality and mortality for all causes of death from 2003 to 2019 as health outcomes.

To estimate the causal effect of the Land of Fires, we exploit causal inference methods that allow us to retrieve estimates of the average treatment effects by controlling for observed demographics and socio-economic differences that may correlate with treatment status, thus biasing the results. Controlling for a large set of covariates is paramount as it allows us to compare two otherwise similar municipalities, but only because one is situated in TdF and the other is not. Specifically, we employ two commonly used techniques: propensity-score matching and IPW regression adjustment.

Using the IPWRA method, along with a series of sensitivity analyses and robustness checks, we show that individuals residing in TdF experience higher cancer mortality rates than the rest of the region and a counterfactual scenario in which no municipality was affected by hazardous waste disposal.

Our findings align closely with previous studies documenting the adverse health effects on affected populations. Furthermore, the elevated cancer mortality observed in TdF is consistent with international research, indicating that exposure to the disposal and combustion of hazardous waste negatively impacts public health.

However, our data do not allow us to determine whether the effects vary based on geographical proximity to open-air (illegal) landfills or the type and quantity of toxic substances disposed of. Furthermore, we cannot assess potential impacts on health outcomes beyond cancer mortality due to data limitations. Therefore, further analysis is needed.

Our findings support the hypothesis that environmental contamination harms individuals' health. Given the high population density in these areas, the potential costs to the National Health Service could be substantial.

Although ten years have passed since the agreement for TdF was signed, the central government never performed the land decontamination initially dictated in the DL 136/2013 "Terra dei Fuochi" nor implemented other measures to defeat these crimes. Even though local authorities have made numerous attempts to neutralize and defeat such phenomena, illegal dumping and waste burning continue, thus calling for more drastic policy interventions from the central government and public health authorities.

Bibliography

Abadie, A. and Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics*, 10:465–503.

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557.

Abadie, A. and Imbens, G. W. (2009). Matching on the estimated propensity score. Working Paper 15301, National Bureau of Economic Research.

Alberti, P. (2022). The 'land of fires': epidemiological research and public health policy during the waste crisis in campania, italy. *Heliyon*, 8(12):e12331.

Altavista, P., Belli, S., Bianchi, F., Binazzi, A., Comba, P., Del Giudice, R., Fazzo, L., Felli, A., Mastrantonio, M., Menegozzo, M., Musmeci, L., Pizzuti, R., Savarese, A., Trinca, S., and Uccelli, R. (2004). Cause-specific mortality in an area of Campania with numerous waste disposal sites. *Epidemiol Prev*, 28(6):311–321.

ARITUM (2015). I numeri del cancro in Italia – 2015.

ARPAC (2005). Piano regionale di bonifica dei siti inquinati della regione Campania.

ARPAC (2017). "Terra dei fuochi": i risultati delle ultime analisi.

Bencivelli, S. (2016). Domande sulla terra dei fuochi. *Le Scienze*, 2:29–37.

Cafieri, S. and Feoli, F. (2023). Ageing and pollution in the "terra dei fuochi". *Rivista Italiana di Economia Demografia e Statistica*, pages 55–66.

Caliendo, M. and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72.

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

De Crescenzo, D. (2021). Terra dei fuochi, bonifiche e riforme ferme da 13 anni. *Lavialibera*.

De Felip, E., Bianchi, F., Bove, C., Cori, L., D'Argenzio, A., D'Orsi, G., Fusco, M., Miniero, R., Ortolani, R., Palombino, R., Parlato, A., Pelliccia, M. G., Peluso, F., Piscopo, G., Pizzuti, R., Porpora, M. G., Protano, D., Senofonte, O., Spena, S. R., Simonetti, A., and di Domenico, A. (2014). Priority persistent contaminants in people dwelling in critical areas of campania region, italy (sebiorec biomonitoring study). *Sci Total Environ*, 487:420–435.

DiPrete, T. A. and Gangl, M. (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology*, 34(1):271–310.

Dolk, H., Vrijheid, M., Armstrong, B., Abramsky, L., Bianchi, F., Garne, E., Nelen, V., Robert, E., Scott, J. E., Stone, D., and Tenconi, R. (1998). Risk of congenital anomalies near hazardous-waste landfill sites in europe: the eurohazcon study. *Lancet*, 352(9126):423–427.

Fazzo, L., De Santis, M., Beccaloni, E., Scaini, F., Iavarone, I., Comba, P., and Airoma, D. (2020). A geographic information system-based indicator of waste risk to investigate the health impact of landfills and uncontrolled dumping sites. *International journal of environmental research and public health*, 17(16):5789.

Fazzo, L., Manno, V., Iavarone, I., Minelli, G., De Santis, M., Beccaloni, E., Scaini, F., Miotto, E., Airoma, D., and Comba, P. (2023). The health impact of hazardous waste landfills and illegal dumps contaminated sites: An epidemiological study at ecological level in italian region. *Frontiers in public health*, 11:996960.

Fazzo, L., Minichilli, F., Santoro, M., Ceccarini, A., Della Seta, M., Bianchi, F., Comba, P., and Martuzzi, M. (2017). Hazardous waste and health impact: a systematic review of the scientific literature. *Environmental Health*, 16(1):107.

Fusco, M., Guida, A., Bidoli, E., Ciullo, V., Vitale, M. F., Savoia, F., Pirozzi, A., and Serraino, D. (2020). Mortality atlas of the campania region: All-cause and cause-specific mortality at municipal level, 2006-2014. *Rivista dell'Associazione italiana di epidemiologia*.

Garcia-Perez, J., Fernandez-Navarro, P., Castello, A., Lopez-Cima, M. F., Ramis, R., Boldo, E., and Lopez-Abente, G. (2013). Cancer mortality in towns in the vicinity of incinerators and installations for the recovery or disposal of hazardous waste. *Environment International*, 51:31–44.

Geschwind, S. A., Stolwijk, J. A., Bracken, M., Fitzgerald, E., Stark, A., Olsen, C., and Melius, J. (1992). Risk of congenital malformations associated with proximity to hazardous waste sites. *Am J Epidemiol*, 135(11):1197–1207.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

ISS (2020). Studio sull'impatto sanitario degli smaltimenti controllati ed abusivi di rifiuti nei 38 comuni del circondario della procura della Repubblica di Napoli Nord – Rapporto Finale.

Istat (2018). Codifica cause di morte con ICD-10, versione 2016: Istruzioni adottate dall'Istat e orientate al sistema Iris.

Istat (2020). Dati statistici per il territorio – Regione Campania.

Jarup, L., Briggs, D., de Hoogh, C., Morris, S., Hurt, C., Lewin, A., Maitland, I., Richardson, S., Wakefield, J., and Elliott, P. (2002). Cancer risks in populations living near landfill sites in great britain. *Br J Cancer*, 86(11):1732–1736.

Lauria, L. and Spinelli, Angela Trinca, S. (2004). Studio esplorativo su esiti riproduttivi in are con presenza di discariche. In Rapporti ISTISAN, editor, *Valutazione del rischio sanitario e ambientale nello smaltimento di rifiuti urbani e pericolosi*, pages 105–122. ISTISAN.

Legambiente (2003). Rapporto Ecomafia2003.

Legambiente (2013). Terra dei fuochi: radiografia di un ecocidio.

Maselli, V., Polese, G., Rippa, D., Ligrone, R., Kumar Rastogi, R., and Fulgione, D. (2010). Frogs, sentinels of dna damage induced by pollution in naples and the neighbouring provinces. *Ecotoxicol Environ Saf*, 73(7):1525–1529.

Mazza, A., Piscitelli, P., Falco, A., Santoro, M. L., Colangelo, M., Imbriani, G., Idolo, A., De Donno, A., Iannuzzi, L., and Colao, A. (2018). Heavy environmental pressure in campania and other italian regions: A short review of available evidence. *Int J Environ Res Public Health*, 15(1).

Mazza, A., Piscitelli, P., Neglia, C., Rosa, G. D., and Iannuzzi, L. (2015). Illegal dumping of toxic waste and its effect on human health in campania, italy. *International journal of environmental research and public health*, 12(6):6818–6831.

Ministero dell'Interno (2018). Il fenomeno degli incendi dolosi di rifiuti nella "Terra dei Fuochi". Strumenti e modalità di contrasto.

Musmeci, L., Comba, P., Fazzo, L., Iavarone, I., Salmaso, S., Conti, S., Manno, V., and Minelli, G. (2015). Mortality, hospitalization and cancer incidence in the "Land of Fires" in Campania (report required by the Law 6/2014). *Rapporto ISTISAN*.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. *Ann. Agricultural Sciences*, pages 1–51.

NVVIP (2013). Il contesto socio-sanitario in Campania: Elementi di analisi territoriale.

Pasetto, R., Zengarin, N., Caranci, N., De Santis, M., Minichilli, F., Santoro, M., Pirastu, R., and Comba, P. (2017). Environmental justice nel sistema di sorveglianza epidemiologica sentieri. *Epidemiol Prev*, 41(2):134–139.

Pirastu, R., Pasetto, R., Zona, A., Ancona, C., Iavarone, I., Martuzzi, M., and Comba, P. (2013). The health profile of populations living in contaminated sites: Sentieri approach. *Journal of Environmental and Public Health*, 2013(1):939267.

Ravindra, K., Chanana, N., and Mor, S. (2021). Exposure to air pollutants and risk of congenital anomalies: A systematic review and metaanalysis. *Sci Total Environ*, 765:142772.

Rivezzi, G., Piscitelli, P., Scortichini, G., Giovannini, A., Diletti, G., Migliorati, G., Ceci, R., Rivezzi, G., Cirasino, L., Carideo, P., Black, D. M., Garzillo, C., and Giani, U. (2013). A general model of dioxin contamination in breast milk: results from a study on 94 women from the caserta and naples areas in italy. *Int J Environ Res Public Health*, 10(11):5953–5970.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121.

Rocco, G., Petitti, T., Martucci, N., Piccirillo, M. C., La Rocca, A., La Manna, C., De Luca, G., Morabito, A., Chirico, A., Franco, R., et al. (2016). Survival after surgical treatment of lung cancer arising in the population exposed to illegal dumping of toxic waste in the land of fires ('Terra dei Fuochi') of Southern Italy. *Anticancer research*, 36(5):2119–2124.

Rosenbaum, P. R. (2002). *Observational Studies*, pages 1–17. Springer New York, New York, NY.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Saviano, R. (2006). *Gomorra: Viaggio nell'impero economico e nel sogno di dominio della camorra*. Mondadori.

Selvaraj, R. C., Cioffi, G., Waite, K. A., Jackson, S. S., and Barnholtz-Sloan, J. S. (2025). A pan-cancer analysis of age and sex differences in cancer incidence and survival in the united states, 2001-2020. *Cancers*, 17(3).

Senior, K. and Mazza, A. (2004). Italian "Triangle of death" linked to waste crisis. *The Lancet Oncology*, 5(9):525–527.

Svimez (2022). Promozione di iniziative di studio e ricerca socioeconomica a supporto della programmazione e valutazione delle politiche di coesione della regione Campania – Dinamiche demografiche e percorsi di sviluppo territoriale.

Tortora, F. (1989). Affare rifiuti, il pci spiega come sottrarlo alla camorra. *Il Giornale di Napoli*.

Trinca, S., Comba, P., Felli, A., Forte, T., Musmeci, L., and Piccardi, A. (2001). Childhood mortality in an area of southern italy with numerous dumping grounds: application of gis and preliminary findings. In *European Conference Geographic Information Sciences in Public Health: 2001; Sheffield*, volume 19.

Vazquez-Bare, G. (2023). Identification and estimation of spillover effects in randomized experiments. *Journal of Econometrics*, 237(1):105237.

WHO (2024). *Chapter 4. Solid waste. In: Compendium of WHO and other UN guidance in health and environment, 2024 update*. World Health Organization.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141(2):1281–1301.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Appendix A – Additional Results

Table A.1: Descriptive regressions on cancer-specific mortality rate – Results for males.

	(1)	(2)	(3)	(4)	(5)
Tdf=1	-6.10***	1.84***	1.66***	1.66***	0.65
	(0.56)	(0.59)	(0.58)	(0.59)	(0.68)
	(0.5616)	(0.5929)	(0.5845)	(0.5850)	(0.6834)
Constant	31.95***	40.77***	26.77***	24.31***	16.04
	(0.37)	(9.61)	(9.23)	(9.25)	(10.13)
Pop. & Death (in levels)	No	Yes	Yes	Yes	Yes
Pop. & Death (distr. by age)	No	No	Yes	Yes	Yes
X	No	No	No	No	Yes
Year FE	No	No	No	Yes	Yes
Adj. R-squared	0.01	0.06	0.07	0.07	0.07
Observations	9265	9265	9265	9265	9265
F	118.07	61.78	44.29	21.24	19.03
p	0.00	0.00	0.00	0.00	0.00

Notes: Coefficients are reported with cluster standard error in parentheses. Standard errors were clustered at the municipality level to account for the potential serial correlation of the error term. The outcome of interest is the cancer-specific mortality rate. **X** is a vector including altitude zone, altitude (in meters), a dummy indicating whether it is a coastal town, coastal area, level of urbanization, income per capita, and population density. ***, **, and * denote significance at 1%, 5%, and 10%, respectively.

Table A.2: Descriptive regressions on cancer-specific mortality rate – Results for females.

	(1)	(2)	(3)	(4)	(5)
Tdf=1	-2.98***	0.89**	0.91*	0.91*	1.00*
	(0.43)	(0.45)	(0.52)	(0.52)	(0.56)
Constant	19.49***	14.14**	10.90*	8.85	9.58
	(0.23)	(6.04)	(6.46)	(6.42)	(6.74)
Pop. & Death (in levels)	No	Yes	Yes	Yes	Yes
Pop. & Death (distr. by age)	No	No	Yes	Yes	Yes
X	No	No	No	No	Yes
Year FE	No	No	No	Yes	Yes
Adj. R-squared	0.01	0.04	0.04	0.05	0.05
Observations	9265	9265	9265	9265	9265
F	48.21	37.45	32.71	16.31	15.44
p	0.00	0.00	0.00	0.00	0.00

Notes: Coefficients are reported with cluster standard error in parentheses. Standard errors were clustered at the municipality level to account for the potential serial correlation of the error term. The outcome of interest is the cancer-specific mortality rate. **X** is a vector including altitude zone, altitude (in meters), a dummy indicating whether it is a coastal town, coastal area, level of urbanization, income per capita, and population density. ***, **, and * denote significance at 1%, 5%, and 10%, respectively.

Table A.3: Placebo exercise – untreated municipalities from Naples & Caserta

	Males				Females			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
	$Pr = \frac{1}{3}$	$Pr = \frac{1}{4}$	$Pr = \frac{1}{5}$	$Pr = \frac{1}{6}$	$Pr = \frac{1}{3}$	$Pr = \frac{1}{4}$	$Pr = \frac{1}{5}$	$Pr = \frac{1}{6}$
ATE (TdF vs non TdF)	0.61	-1.13	2.34**	-1.20	0.25	1.08	0.60	-1.29
	(0.02)	(0.29)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)
ATT	0.39	-1.11	1.92*	-1.12	0.44	1.20	1.11	-1.33
	(0.01)	(0.31)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)
Pop. & Death (in levels)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Pop. & Death (distr. by age)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
X	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,530	1,530	1,530	1,530	1,530	1,530	1,530	1,530

Notes: This table presents the results of the placebo analysis based on $1,000$ permutations. Specifically, it reports the t-statistics for the ATE and ATT obtained from the permutations, along with standard errors (in parentheses) for a two-sided test. The outcome of interest is the cancer-specific mortality rate. Estimates are performed via IPWRA. *Pop.* stands for population and *distr* for distribution. $\mathbf{X}$ is a vector including altitude zone, altitude (in meters), a dummy indicating whether it is a coastal town, coastal area, level of urbanization, income per capita, and population density. ***, **, and * denote significance at 1%, 5%, and 10%, respectively.

Table A.4: IPWRA results for cancer-specific mortality rate – Naples & Caserta

	Males			Females		
	(1)	(2)	(3)	(1)	(2)	(3)
ATE (TdF vs non TdF)	1.47*	2.27***	1.55**	3.61***	4.98***	3.95***
	(0.75)	(0.71)	(0.67)	(0.60)	(0.71)	(0.70)
ATT	0.65	5.66***	4.32***	1.31*	5.11***	3.92***
	(0.76)	(1.51)	(1.47)	(0.68)	(1.13)	(0.91)
Pop. & Death (in levels)	Yes	Yes	Yes	Yes	Yes	Yes
Pop. & Death (distr. by age)	Yes	Yes	Yes	Yes	Yes	Yes
X	No	Yes	Yes	No	Yes	Yes
Year FE	No	No	Yes	No	No	Yes
Observations	3,196	3,295	3,296	3,298	3,257	3,226

Notes: Coefficients are reported with cluster-robust robust standard errors in parentheses clustered at the municipality level to account for the potential serial correlation in the mortality rate. The outcome of interest is the cancer-specific mortality rate. Estimates are performed via IPWRA. *Pop.* stands for population and *distr* for distribution. $\mathbf{X}$ is a vector including altitude zone, altitude (in meters), a dummy indicating whether it is a coastal town, coastal area, level of urbanization, income per capita, and population density. ***, **, and * denote significance at 1%, 5%, and 10%, respectively.

Table A.5: IPWRA results for cancer-specific mortality rate – excluding untreated municipalities at the border

	Males			**Females**		
	(1)	(2)	(3)	(1)	(2)	(3)
ATE (TdF vs non TdF)	3.11***	2.11***	1.63**	4.44***	7.30***	6.27***
ATT	2.13**	4.26***	3.08***	1.80***	3.68***	3.10***
	(0.92)	(0.89)	(0.81)	(0.54)	(0.91)	(0.74)
Pop. & Death (in levels)	Yes	Yes	Yes	Yes	Yes	Yes
Pop. & Death (distr. by age)	Yes	Yes	Yes	Yes	Yes	Yes
X	No	Yes	Yes	No	Yes	Yes
Year FE	No	No	Yes	No	No	Yes
Observations	4,335	6,761	6,815	4,913	6,680	6,739

Notes: Coefficients are reported with cluster-robust robust standard errors in parentheses clustered at the municipality level to account for the potential serial correlation in the mortality rate. The outcome of interest is the cancer-specific mortality rate. Estimates are performed via IPWRA. *Pop.* stands for population and *distr* for distribution. **X** is a vector including altitude zone, altitude (in meters), a dummy indicating whether it is a coastal town, coastal area, level of urbanization, income per capita, and population density. ***, **, and * denote significance at 1%, 5%, and 10%, respectively.

Table A.6: IPWRA results for cancer-specific mortality rate – Pre vs Post Introduction DL TdF

	Males			**Females**		
	(1)	(2)	(3)	(1)	(2)	(3)
Panel (a): Pre Introduction DL "Terra dei Fuochi"						
ATE (TdF vs non TdF)	2.45***	1.48*	1.00	7.50***	5.20***	5.39***
	(0.54)	(0.88)	(0.81)	(0.58)	(0.66)	(0.65)
ATT	2.53***	5.08***	4.78***	1.52**	3.30***	3.18***
	(0.96)	(0.87)	(0.89)	(0.62)	(0.77)	(0.73)
Pop. & Death (in levels)	Yes	Yes	Yes	Yes	Yes	Yes
Pop. & Death (distr. by age)	Yes	Yes	Yes	Yes	Yes	Yes
X	No	Yes	Yes	No	Yes	Yes
Year FE	No	No	Yes	No	No	Yes
Observations	3,938	5,002	5,020	5,148	4,866	4,891

	Males			**Females**		
	(1)	(2)	(3)	(1)	(2)	(3)
Panel (b): Post Introduction DL "Terra dei Fuochi"						
ATE (TdF vs non TdF)	6.19***	2.61**	2.70**	5.40***	7.89***	7.81***
	(1.03)	(1.24)	(1.25)	(1.24)	(0.95)	(0.81)
ATT	1.82	0.73	0.53	2.31***	3.26***	2.94***
	(1.48)	(1.30)	(1.35)	(0.43)	(0.67)	(0.63)
Pop. & Death (in levels)	Yes	Yes	Yes	Yes	Yes	Yes
Pop. & Death (distr. by age)	Yes	Yes	Yes	Yes	Yes	Yes
X	No	Yes	Yes	No	Yes	Yes
Year FE	No	No	Yes	No	No	Yes
Observations	2,430	2,774	2,789	2,658	2,691	2,696

Notes: Coefficients are reported with cluster-robust robust standard errors in parentheses clustered at the municipality level to account for the potential serial correlation in the mortality rate. The outcome of interest is the cancer-specific mortality rate. Estimates are performed via IPWRA. *Pop.* stands for population and *distr* for distribution. **X** is a vector including altitude zone, altitude (in meters), a dummy indicating whether it is a coastal town, coastal area, level of urbanization, income per capita, and population density. ***, **, and * denote significance at 1%, 5%, and 10%, respectively.

Table A.7: Propensity-score Matching results for cancer-specific mortality rate

	Males			Females		
	(1)	(2)	(3)	(1)	(2)	(3)
ATE (TdF vs non TdF)	-2.52	-7.43**	-8.21**	6.79**	2.54***	5.03***
	(2.27)	(2.99)	(3.56)	(2.77)	(0.83)	(0.97)
ATT	-1.08	2.31	4.41***	0.70	2.34*	4.76***
	(0.72)	(2.46)	(1.58)	(0.59)	(1.31)	(1.75)
Pop. & Death (in levels)	Yes	Yes	Yes	Yes	Yes	Yes
Pop. & Death (distr. by age)	Yes	Yes	Yes	Yes	Yes	Yes
X	No	Yes	Yes	No	Yes	Yes
Year FE	No	No	Yes	No	No	Yes
Observations	7,837	7,743	7,806	7,956	7,515	7,590
Matches: requested	1	1	1	1	1	1
Min matches	17	1	1	17	1	1
Max matches	17	1	1	17	1	1

Notes: Coefficients are reported with Abadie-Imbens heteroskedastic robust standard errors in parentheses (Abadie and Imbens, 2006). The outcome of interest is the cancer-specific mortality rate. Estimates are performed via nearest neighbor propensity-score matching, where estimates of the propensity score are obtained via logit. *Pop.* stands for population and *distr* for distribution. **X** is a vector including altitude zone, altitude (in meters), a dummy indicating whether it is a coastal town, coastal area, level of urbanization, income per capita, and population density. ***, **, and * denote significance at 1%, 5%, and 10%, respectively.

Table A.8: Sensitivity Propensity-score Matching – varying number of neighbors

	Males				Females			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
ATE (TdF vs non TdF)	-8.21**	-4.89**	-3.85**	-4.18***	5.03***	4.01***	8.69*	5.93*
	(3.56)	(2.49)	(1.69)	(1.20)	(0.97)	(0.68)	(4.74)	(3.59)
ATT	4.41***	3.57***	2.95***	2.84***	4.76***	3.38***	3.04***	3.23***
	(1.58)	(0.97)	(0.72)	(0.58)	(1.75)	(0.89)	(0.60)	(0.44)
Pop. & Death (in levels)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Pop. & Death (distr. by age)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
X	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	7,806	7,806	7,806	7,806	7,590	7,590	7,590	7,590
Matches requested	1	2	3	4	1	2	3	4
Min matches	1	2	3	4	1	2	3	4
Max matches	1	2	3	4	1	2	3	4

Notes: Coefficients are reported with Abadie-Imbens heteroskedastic robust standard errors in parentheses (Abadie and Imbens, 2006). The outcome of interest is the cancer-specific mortality rate. Estimates are performed via nearest neighbor propensity-score matching, where estimates of the propensity score are obtained via logit. *Pop.* stands for population and *distr* for distribution. **X** is a vector including altitude zone, altitude (in meters), a dummy indicating whether it is a coastal town, coastal area, level of urbanization, income per capita, and population density. ***, **, and * denote significance at 1%, 5%, and 10%, respectively.

Table A.9: Sensitivity Propensity-score Matching – varying caliper

	Males				Females			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
ATE (TdF vs non TdF)	-8.21**	-8.21**	-8.21**	-8.21**	5.03***	5.03***	5.03***	5.03***
	(3.56)	(3.56)	(3.56)	(3.56)	(0.97)	(0.97)	(0.97)	(0.97)
ATT	4.41***	4.41***	4.41***	4.41***	4.76***	4.76***	4.76***	4.76***
	(1.58)	(1.58)	(1.58)	(1.58)	(1.75)	(1.75)	(1.75)	(1.75)
Pop. & Death (in levels)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Pop. & Death (distr. by age)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
X	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	7,806	7,806	7,806	7,806	7,590	7,590	7,590	7,590
Matches requested	1	1	1	1	1	1	1	1
Min matches	1	1	1	1	1	1	1	1
Max matches	1	1	1	1	1	1	1	1

Notes: Coefficients are reported with Abadie-Imbens heteroskedastic robust standard errors in parentheses (Abadie and Imbens, 2006). The outcome of interest is the cancer-specific mortality rate. Estimates are performed via nearest neighbor propensity-score matching, where estimates of the propensity score are obtained via logit. *Pop.* stands for population and *distr* for distribution. **X** is a vector including altitude zone, altitude (in meters), a dummy indicating whether it is a coastal town, coastal area, level of urbanization, income per capita, and population density. ***, **, and * denote significance at 1%, 5%, and 10%, respectively.

Table A.10: Different Propensity-score matching algorithms

	Males								Females							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	one-to-one	one-to-one no over.	No rep.	Kernel	LLR	Radius	Common	Trim	one-to-one	one-to-one no over.	No rep.	Kernel	LLR	Radius	Common	Trim
ATT	4.41***	4.38***	-2.06***	2.39*	1.14	-6.10***	0.98	1.26*	4.76***	4.70***	-0.26	2.42***	1.74**	-2.98***	0.74*	2.05***
	(1.61)	(1.61)	(0.35)	(1.34)	(1.34)	(0.32)	(1.34)	(0.73)	(1.78)	(1.78)	(0.27)	(0.79)	(0.72)	(0.26)	(0.45)	(0.53)
Pop. & Death (in levels)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Pop. & Death (distr. by age)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
X	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	7,806	9,265	9,265	9,265	9,265	9,265	9,265	9,265	7,590	9,265	9,265	9,265	9,265	9,265	9,265	9,265

Notes: The outcome of interest is the cancer-specific mortality rate. Estimates are performed via nearest neighbor propensity-score matching, where estimates of the propensity score are obtained via logit. *Pop.* stands for population and *distr* for distribution. **X** is a vector including altitude zone, altitude (in meters), a dummy indicating whether it is a coastal town, coastal area, level of urbanization, income per capita, and population density. ***, **, and * denote significance at 1%, 5%, and 10%, respectively.

Table A.11: Rosenbaum Bounds

	Γ	sig+	sig-	t-hat+	t-hat-	CI+	CI-
Panel (a): Males							
	1.00	0	0	4.77488	4.77488	4.15416	5.39215
	1.25	0	0	3.58915	5.94626	2.95463	6.54959
	1.5	1.90e-14	0	2.61313	6.87306	1.96943	7.47318
	1.75	1.10e-07	0	1.7924	7.63813	1.13649	8.2366
	2.00	0.000886	0	1.08633	8.28334	0.41466	8.88393
	2.25	0.091489	0	0.46164	8.84033	-0.221604	9.44602
	2.5	0.606189	0	-0.092675	9.3326	-0.796624	9.94044
	2.75	0.957123	0	-0.595225	9.76944	-1.30833	10.3757
	3.00	0.998834	0	-1.05153	10.1593	-1.7804	10.7723
Panel (b): Females							
	1.00	0	0	4.76508	4.76508	4.25713	5.27071
	1.25	0	0	3.79309	5.73277	3.27825	6.24615
	1.5	0	0	3.00635	6.52107	2.48364	7.03811
	1.75	3.30e-16	0	2.34184	7.17824	1.81136	7.69661
	2.00	5.80e-10	0	1.77023	7.73814	1.22044	8.26592
	2.25	6.70e-06	0	1.25954	8.22897	0.708684	8.75947
	2.5	0.002418	0	0.810856	8.65813	0.255044	9.19194
	2.75	0.075571	0	0.415073	9.04093	-0.156045	9.5777
	3.00	0.429936	0	0.052122	9.38483	-0.5335	9.92255

Notes: Γ stands for the log odds of differential assignment due to unobserved factors; *sig+* for the upper bound significance level; *sig-* for the lower bound significance level; *t-hat+* for the upper bound Hodges-Lehmann point estimate; *t-hat-* for the lower bound Hodges-Lehmann point estimate; *CI+* for the upper bound for the 95% confidence interval; and *CI-* for the lower bound for the 95% confidence interval.

Bibliography

Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies*, 72(1):1–19.

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425.

Abadie, A. and Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics*, 10:465–503.

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. *American economic review*, 93(1):113–132.

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557.

Abadie, A. and Imbens, G. W. (2009). Matching on the estimated propensity score. Working Paper 15301, National Bureau of Economic Research.

Aimone Gigio, L., Alampi, D., Camussi, S., Ciaccio, G., Guaitini, P., Lozzi, M., Mancini, A. L., Panicara, E., and Paolicelli, M. (2018). La sanità in Italia: il difficile equilibrio tra vincoli di bilancio e qualità dei servizi nelle regioni in piano di rientro. *Questioni di Economia e Finanza in Bank of Italy*, 427:1–100.

Alberti, P. (2022). The 'land of fires': epidemiological research and public health policy during the waste crisis in campania, italy. *Heliyon*, 8(12):e12331.

Altavista, P., Belli, S., Bianchi, F., Binazzi, A., Comba, P., Del Giudice, R., Fazzo, L., Felli, A., Mastrantonio, M., Menegozzo, M., Musmeci, L., Pizzuti, R., Savarese, A., Trinca, S., and Uccelli, R. (2004). Cause-specific mortality in an area of Campania with numerous waste disposal sites. *Epidemiol Prev*, 28(6):311–321.

Amarilla-Donoso, F. J., López-Espuela, F., Roncero-Martín, R., Leal-Hernandez, O., Puerto-Parejo, L. M., Aliaga-Vera, I., Toribio-Felipe, R., and Lavado-García, J. M. (2020). Quality of life in elderly people after a hip fracture: a prospective study. *Health and quality of life outcomes*, 18:1–10.

Andrews, M., Elamin, O., Hall, A. R., Kyriakoulis, K., and Sutton, M. (2017). Inference in the presence of redundant moment conditions and the impact of government health expenditure on health outcomes in England. *Econometric Reviews*, 36(1-3):23–41.

Angelillo, I., Ricciardi, G., Nante, N., Boccia, A., Bianco, A., La Torre, G., Vinci, V., and De Giusti, M. (2000). Appropriateness of hospital utilisation in Italy. *Public health*, 114(1):9–14.

Arcà, E., Principe, F., and Van Doorslaer, E. (2020). Death by austerity? The impact of cost containment on avoidable mortality in Italy. *Health Economics*, 29(12):1500–1516.

ARITUM (2015). I numeri del cancro in Italia – 2015.

ARPAC (2005). Piano regionale di bonifica dei siti inquinati della regione Campania.

ARPAC (2017). "Terra dei fuochi": i risultati delle ultime analisi.

Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497.

Balia, S., Brau, R., and Marrocu, E. (2018). Interregional patient mobility in a decentralized healthcare system. *Regional Studies*, 52(3):388–402.

Bei, X. (2024). Inference on union bounds with applications to did, rdd, bunching, and structural counterfactuals.

Bencivelli, S. (2016). Domande sulla terra dei fuochi. *Le Scienze*, 2:29–37.

Beraldo, S., Collaro, M., and Marino, I. (2023). Patient migration as a response to the regulation of subnational healthcare budgets. *Regional Studies*, 0(0):1–13.

Berchialla, P., Foltran, F., Di Lenarda, A., Barbati, G., Franchin, L., Salvatore, L., Gregori, D., et al. (2010). Hospitalization rate as a proxy for the quality of life evaluation: the case of dilated cardiomiopathy. *Journal of preventive medicine and hygiene*, 51(4):152–156.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1):249–275.

Blom, M. C., Jonsson, F., Landin-Olsson, M., and Ivarsson, K. (2014). The probability of patients being admitted from the emergency department is negatively correlated to in-hospital bed occupancy–a registry study. *International journal of emergency medicine*, 7:1–7.

Boden, D., Agarwal, A., Hussain, T., Martin, S., Radford, N., Riyat, M., So, K., Su, Y., Turvey, A., and Whale, C. (2016). Lowering levels of bed occupancy is associated with decreased inhospital mortality and improved performance on the 4-hour target in a UK District General Hospital. *Emergency Medicine Journal*, 33(2):85–90.

Bonhomme, S. and Sauder, U. (2011). Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling. *The Review of Economics and Statistics*, 93(2):479–494.

Bordignon, M., Coretti, S., Piacenza, M., and Turati, G. (2020). Hardening subnational budget constraints via administrative subordination: The Italian experience of recovery plans in regional health services. *Health Economics*, 29(11):1378–1399.

Borra, C., Pons-Pons, J., and Vilar-Rodriguez, M. (2020). Austerity, healthcare provision, and health outcomes in Spain. *The European Journal of Health Economics*, 21:409–423.

Borusyak, K., Jaravel, X., and Spiess, J. (2021). Revisiting event study designs: Robust and efficient estimation. *arXiv preprint:2108.12419*.

Brown, J. R., Chang, C.-H., Zhou, W., MacKenzie, T. A., Malenka, D. J., and Goodman, D. C. (2014). Health system characteristics and rates of readmission after acute myocardial infarction in the United States. *Journal of the American Heart Association*, 3(3):e000714.

Butts, K. (2024). Difference-in-Differences Estimation with Spatial Spillovers. Papers 2105.03737, arXiv.org.

Cafieri, S. and Feoli, F. (2023). Ageing and pollution in the "terra dei fuochi". *Rivista Italiana di Economia Demografia e Statistica*, pages 55–66.

Calabrò, R. (2016). Servono terapie innovative per le regioni in piano di rientro (in Italian). *Il Sole 24 ore, 4 Oct.*

Caliendo, M. and Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72.

Callaway, B. and Li, T. (2019). Quantile treatment effects in difference in differences models with panel data. *Quantitative Economics*, 10(4):1579–1618.

Callaway, B., Li, T., and Oka, T. (2018). Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods. *Journal of Econometrics*, 206(2):395–413. Special issue on Advances in Econometric Theory: Essays in honor of Takeshi Amemiya.

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

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

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources*, 50(2):317–372.

Carneiro, P., Hansen, K. T., and Heckman, J. (2003). Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. IZA Discussion Papers 767, Institute of Labor Economics (IZA).

Chernozhukov, V., Fernandez-Val, I., and Kaji, T. (2016). Extremal quantile regression: An overview. *Handbook of Quantile Regression*.

Chernozhukov, V., Fernandez-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268.

Ciaccio, A. (2023). The impact of a cost-containment measure on the quality of regional health services in italy: a parametric and a non-parametric approach. Working Papers 2023: 24, Department of Economics, University of Venice "Ca' Foscari".

Cirulli, V. and Marini, G. (2023). Are austerity measures really distressing? Evidence from Italy. *Economics & Human Biology*, page 101217.

Crémieux, P., Ouellette, P., and Pilon, C. (1999). Health care spending as determinants of health outcomes. *Health Economics*, 8(7):627–639.

Davidson, R., MacKinnon, J. G., et al. (2004). *Econometric theory and methods*, volume 5. Oxford University Press New York.

De Chaisemartin, C. and d'Haultfoeuille, X. (2022a). Difference-in-differences estimators of intertemporal treatment effects. Technical report, National Bureau of Economic Research.

De Chaisemartin, C. and d'Haultfoeuille, X. (2022b). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. Technical report, National Bureau of Economic Research.

De Chaisemartin, C. and d'Haultfoeuille, X. (2023). Two-way fixed effects and differences-in-differences estimators with several treatments. *Journal of Econometrics*, 236(2):105480.

De Crescenzo, D. (2021). Terra dei fuochi, bonifiche e riforme ferme da 13 anni. *Lavialibera*.

De Felip, E., Bianchi, F., Bove, C., Cori, L., D'Argenzio, A., D'Orsi, G., Fusco, M., Miniero, R., Ortolani, R., Palombino, R., Parlato, A., Pelliccia, M. G., Peluso, F., Piscopo, G., Pizzuti, R., Porpora, M. G., Protano, D., Senofonte, O., Spena, S. R., Simonetti, A., and di Domenico, A. (2014). Priority persistent contaminants in people dwelling in critical areas of campania region, italy (sebiorec biomonitoring study). *Sci Total Environ*, 487:420–435.

De Luca, G., Lisi, D., Martorana, M., and Siciliani, L. (2021). Does higher institutional quality improve the appropriateness of healthcare provision? *Journal of Public Economics*, 194:104356.

Depalo, D. (2019). The side effects on health of a recovery plan in Italy: A nonparametric bounding approach. *Regional Science and Urban Economics*, 78:103466.

DiPrete, T. A. and Gangl, M. (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology*, 34(1):271–310.

Dolk, H., Vrijheid, M., Armstrong, B., Abramsky, L., Bianchi, F., Garne, E., Nelen, V., Robert, E., Scott, J. E., Stone, D., and Tenconi, R. (1998). Risk of congenital anomalies near hazardous-waste landfill sites in europe: the eurohazcon study. *Lancet*, 352(9126):423–427.

Eurostat (2013). European system of accounts - ESA 2010. Technical report, Eurostat.

Fan, Y. and Yu, Z. (2012). Partial identification of distributional and quantile treatment effects in difference-in-differences models. *Economics Letters*, 115(3):511–515.

Fazzo, L., De Santis, M., Beccaloni, E., Scaini, F., Iavarone, I., Comba, P., and Airoma, D. (2020). A geographic information system-based indicator of waste risk to investigate the health impact of landfills and uncontrolled dumping sites. *International journal of environmental research and public health*, 17(16):5789.

Fazzo, L., Manno, V., Iavarone, I., Minelli, G., De Santis, M., Beccaloni, E., Scaini, F., Miotto, E., Airoma, D., and Comba, P. (2023). The health impact of hazardous waste landfills and illegal dumps contaminated sites: An epidemiological study at ecological level in italian region. *Frontiers in public health*, 11:996960.

Fazzo, L., Minichilli, F., Santoro, M., Ceccarini, A., Della Seta, M., Bianchi, F., Comba, P., and Martuzzi, M. (2017). Hazardous waste and health impact: a systematic review of the scientific literature. *Environmental Health*, 16(1):107.

Ferman, B. and Pinto, C. (2016). Revisiting the synthetic control estimator.

Fiorini, M., Lee, W., and Pfeifer, G. (2024). A simple approach to staggered difference-in-differences in the presence of spillovers. *CESifo Working Paper No. 11011*.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1):259–276.

Franklin, B., Hochlaf, D., and Holley-Moore, G. (2017). Public Health in Europe during the austerity years. In *A research report from ICL-UK*. ICL-IK London.

Fusco, M., Guida, A., Bidoli, E., Ciullo, V., Vitale, M. F., Savoia, F., Pirozzi, A., and Serraino, D. (2020). Mortality atlas of the campania region: All-cause and cause-specific mortality at municipal level, 2006-2014. *Rivista dell'Associazione italiana di epidemiologia*.

Garcia-Perez, J., Fernandez-Navarro, P., Castello, A., Lopez-Cima, M. F., Ramis, R., Boldo, E., and Lopez-Abente, G. (2013). Cancer mortality in towns in the vicinity of incinerators and installations for the recovery or disposal of hazardous waste. *Environment International*, 51:31–44.

Geschwind, S. A., Stolwijk, J. A., Bracken, M., Fitzgerald, E., Stark, A., Olsen, C., and Melius, J. (1992). Risk of congenital malformations associated with proximity to hazardous waste sites. *Am J Epidemiol*, 135(11):1197–1207.

Ghislandi, S., Renner, A.-T., and Varghese, N. (2025). The impact of budget cuts on individual patient health: Causal evidence from hospital closures. *Journal of Health Economics*, page 102975.

Giancotti, M., Sulku, S. N., Pipitone, V., Mauro, M., et al. (2020). Do Recovery Plans Improve Public Hospitals Efficiency and Productivity? Evidence from Italy. *International Review of Business Research Papers*, 16(1).

Golinelli, D., Toscano, F., Bucci, A., Lenzi, J., Fantini, M. P., Nante, N., and Messina, G. (2017). Health expenditure and all-cause mortality in the 'galaxy' of Italian regional healthcare systems: a 15-year panel data analysis. *Applied health economics and health policy*, 15:773–783.

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

Gregory, K. D., Jackson, S., Korst, L., and Fridman, M. (2012). Cesarean versus vaginal delivery: whose risks? Whose benefits? *American journal of perinatology*, 29(01):07–18.

Guccio, C., Pignataro, G., Romeo, D., and Vidoli, F. (2022). Is austerity good for efficiency, at least? A counterfactual assessment for the Italian NHS. Health, Econometrics and Data Group (HEDG) Working Papers 22/28, HEDG, c/o Department of Economics, University of York.

Guccio, C., Pignatora, G., and Vidoli, F. (2023). It never rains but it pours: Austerity and mortality rate in peripheral areas. Health, Econometrics and Data Group (HEDG) Working Papers 23/02, HEDG, c/o Department of Economics, University of York.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, 66(5):1017–1098.

Heckman, J. J., Humphries, J. E., and Veramendi, G. (2016). Dynamic treatment effects. *Journal of Econometrics*, 191(2):276–292. Innovations in Measurement in Economics and Econometrics.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654.

Heijink, R., Koolman, X., and Westert, G. (2013). Spending more money, saving more lives? The relationship between avoidable mortality and healthcare spending in 14 countries. *The European Journal of Health Economics*, 14(3):527–538.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Hunt, D. J. and Link, C. R. (2020). Better outcomes at lower costs? the effect of public health expenditures on hospital efficiency. *Applied Economics*, 52(4):400–414.

Imai, K. and Kim, I. S. (2021). On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data. *Political Analysis*, 29(3):405–415.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86.

Institute of Medicine (2001). *Crossing the Quality Chasm: A New Health System for the 21st Century*. The National Academies Press, Washington, DC.

ISS (2020). Studio sull'impatto sanitario degli smaltimenti controllati ed abusivi di rifiuti nei 38 comuni del circondario della procura della Repubblica di Napoli Nord – Rapporto Finale.

Istat (2018). Codifica cause di morte con ICD-10, versione 2016: Istruzioni adottate dall'Istat e orientate al sistema Iris.

Istat (2020). Dati statistici per il territorio – Regione Campania.

Jarup, L., Briggs, D., de Hoogh, C., Morris, S., Hurt, C., Lewin, A., Maitland, I., Richardson, S., Wakefield, J., and Elliott, P. (2002). Cancer risks in populations living near landfill sites in great britain. *Br J Cancer*, 86(11):1732–1736.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. CRC press.

Kwok, C. S., Rao, S. V., Potts, J. E., Kontopantelis, E., Rashid, M., Kinnaird, T., Curzen, N., Nolan, J., Bagur, R., and Mamas, M. A. (2018). Burden of 30-day readmissions after percutaneous coronary intervention in 833,344 patients in the United States: predictors, causes, and cost: insights from the Nationwide Readmission Database. *JACC: Cardiovascular Interventions*, 11(7):665–674.

Lauria, L. and Spinelli, Angela Trinca, S. (2004). Studio esplorativo su esiti riproduttivi in are con presenza di discariche. In Rapporti ISTISAN, editor, *Valutazione del rischio sanitario e ambientale nello smaltimento di rifiuti urbani e pericolosi*, pages 105–122. ISTISAN.

Legambiente (2003). Rapporto Ecomafia2003.

Legambiente (2013). Terra dei fuochi: radiografia di un ecocidio.

Leive, A., David, G., and Candon, M. (2023). On resource allocation in health care: The case of concierge medicine. *Journal of Health Economics*, page 102776.

Li, X. and Lin, L. (2024). Identification of quantile treatment effects in difference-in-differences settings with staggered adoption. *Economics Letters*, 241:111792.

Linton, O., Maasoumi, E., and Whang, Y.-J. (2005). Consistent testing for stochastic dominance under general sampling schemes. *The Review of Economic Studies*, 72(3):735–765.

Lobo, M. F., Azzone, V., Lopes, F., Freitas, A., Costa-Pereira, A., Normand, S.-L., and Teixeira-Pinto, A. (2020). Understanding the large heterogeneity in hospital readmissions and mortality for acute myocardial infarction. *Health Policy*, 124(7):684–694.

Maasoumi, E. (2003). *Parametric and Nonparametric Tests of Limited Domain and Ordered Hypotheses in Economics*, chapter 25, pages 538–556. John Wiley & Sons, Ltd.

Maasoumi, E. and Wang, L. (2019). The gender gap between earnings distributions. *Journal of Political Economy*, 127(5):2438–2504.

Machado, J. A. and Santos Silva, J. (2019). Quantiles via moments. *Journal of Econometrics*, 213(1):145–173.

MacKinnon, J. G., Nielsen, M. Ø., and Webb, M. D. (2023). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics*, 232(2):272–299.

MacKinnon, J. G. and Webb, M. D. (2017). Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics*, 32(2):233–254.

MacKinnon, J. G. and Webb, M. D. (2018). The wild bootstrap for few (treated) clusters. *The Econometrics Journal*, 21(2):114–135.

Madsen, F., Ladelund, S., and Linneberg, A. (2014). High levels of bed occupancy associated with increased inpatient and thirty-day hospital mortality in Denmark. *Health affairs*, 33(7):1236–1244.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323.

Manski, C. F. (1994). Analog estimation of econometric models. *Handbook of econometrics*, 4:2559–2582.

Manski, C. F. (2003). *Partial identification of probability distributions*, volume 5. Springer.

Manski, C. F. and Pepper, J. V. (2013). Deterrence and the death penalty: Partial identification analysis using repeated cross sections. *Journal of Quantitative Criminology*, 29:123–141.

Manski, C. F. and Pepper, J. V. (2018). How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions. *The Review of Economics and Statistics*, 100(2):232–244.

Martini, G., Berta, P., Mullahy, J., and Vittadini, G. (2014). The effectiveness–efficiency trade-off in health care: The case of hospitals in Lombardy, Italy. *Regional Science and Urban Economics*, 49:217–231.

Maselli, V., Polese, G., Rippa, D., Ligrone, R., Kumar Rastogi, R., and Fulgione, D. (2010). Frogs, sentinels of dna damage induced by pollution in naples and the neighbouring provinces. *Ecotoxicol Environ Saf*, 73(7):1525–1529.

Mazza, A., Piscitelli, P., Falco, A., Santoro, M. L., Colangelo, M., Imbriani, G., Idolo, A., De Donno, A., Iannuzzi, L., and Colao, A. (2018). Heavy environmental pressure in campania and other italian regions: A short review of available evidence. *Int J Environ Res Public Health*, 15(1).

Mazza, A., Piscitelli, P., Neglia, C., Rosa, G. D., and Iannuzzi, L. (2015). Illegal dumping of toxic waste and its effect on human health in campania, italy. *International journal of environmental research and public health*, 12(6):6818–6831.

McGuire, A. (1987). The measurement of hospital efficiency. *Social Science & Medicine*, 24(9):719–724.

McKay, N. L. and Deily, M. E. (2008). Cost inefficiency and hospital health outcomes. *Health economics*, 17(7):833–848.

McKibbin, R. (2023). The effect of RCTs on drug demand: Evidence from off-label cancer drugs. *Journal of Health Economics*, 90:102779.

Melly, B. and Santangelo, G. (2015). The changes-in-changes model with covariates. *W.P.*

Miller, B. I. (2023). Doubly-robust quantile treatment effect estimation. *W.P.*

Ministero dell'Interno (2018). Il fenomeno degli incendi dolosi di rifiuti nella "Terra dei Fuochi". Strumenti e modalità di contrasto.

MoH (2014). Analisi spazio-temporale degli indicatori della Griglia LEA relativi all'adempimento sul "mantenimento dell'erogazione dei LEA" - Trend regionali anni 2001 - 2012. Technical report, Ministry of Health.

MoH (2015). Monitoraggio dei LEA attraverso la cd. Griglia LEA. Technical report, Ministry of Health.

MoH (2016). Monitoraggio dei LEA attraverso la cd. Griglia LEA. Technical report, Ministry of Health.

MoH (2017). Monitoraggio dei LEA attraverso la cd. Griglia LEA. Technical report, Ministry of Health.

MoH (2018). Monitoraggio dei LEA attraverso la cd. Griglia LEA. Technical report, Ministry of Health.

MoH (2020). Rapporto annuale sull'attività di ricovero ospedaliero - Dati SDO 2020. Technical report, Ministry of Health.

Moja, L., Piatti, A., Pecoraro, V., Ricci, C., Virgili, G., Salanti, G., Germagnoli, L., Liberati, A., and Banfi, G. (2012). Timing matters in hip fracture surgery: patients operated within 48 hours have better outcomes. a meta-analysis and meta-regression of over 190,000 patients. *Public Library of Science San Francisco, USA*.

Musmeci, L., Comba, P., Fazzo, L., Iavarone, I., Salmaso, S., Conti, S., Manno, V., and Minelli, G. (2015). Mortality, hospitalization and cancer incidence in the "Land of Fires" in Campania (report required by the Law 6/2014). *Rapporto ISTISAN*.

Navarro, G., Prat-Marin, A., Asenjo, M., Menacho, A., Trilla, A., and Salleras, L. (2001). Review of the utilisation of a university hospital in Barcelona (Spain): evolution 1992–1996. *European journal of epidemiology*, 17(7):679–684.

Nelsen, R. (2006). *An Introduction to Copulas*. Springer.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. *Ann. Agricultural Sciences*, pages 1–51.

Nixon, J. and Ulmann, P. (2006). The relationship between health care expenditure and health outcomes: evidence and caveats for a causal link. *The European journal of health economics*, 7:7–18.

NVVIP (2013). Il contesto socio-sanitario in Campania: Elementi di analisi territoriale.

OECD (2016). *Scoping paper on health system efficiency measurement (cooperation between the OECD and EC in promoting efficiency in health care)*. OECD.

OECD (2019). *Health at a Glance 2019*. OECD.

OECD (2021). *Health at a Glance 2021*. OECD.

Palmer, S. and Torgerson, D. J. (1999). Definitions of efficiency. *Bmj*, 318(7191):1136.

Pasetto, R., Zengarin, N., Caranci, N., De Santis, M., Minichilli, F., Santoro, M., Pirastu, R., and Comba, P. (2017). Environmental justice nel sistema di sorveglianza epidemiologica sentieri. *Epidemiol Prev*, 41(2):134–139.

Pileggi, C., Bianco, A., Di Stasio, S., and Angelillo, I. (2004). Inappropriate hospital use by patients needing urgent medical attention in Italy. *Public health*, 118(4):284–291.

Pirastu, R., Pasetto, R., Zona, A., Ancona, C., Iavarone, I., Martuzzi, M., and Comba, P. (2013). The health profile of populations living in contaminated sites: Sentieri approach. *Journal of Environmental and Public Health*, 2013(1):939267.

Polidano, C., Zhu, A., and Bornstein, J. C. (2017). The relation between cesarean birth and child cognitive development. *Scientific reports*, 7(1):11483.

Pross, C., Busse, R., and Geissler, A. (2017). Hospital quality variation matters–A time-trend and cross-section analysis of outcomes in German hospitals from 2006 to 2014. *Health policy*, 121(8):842–852.

Quaglio, G., Karapiperis, T., Van Woensel, L., Arnold, E., and McDaid, D. (2013). Austerity and health in Europe. *Health policy*, 113(1-2):13–19.

Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, page rdad018.

Ravindra, K., Chanana, N., and Mor, S. (2021). Exposure to air pollutants and risk of congenital anomalies: A systematic review and metaanalysis. *Sci Total Environ*, 765:142772.

Rivezzi, G., Piscitelli, P., Scortichini, G., Giovannini, A., Diletti, G., Migliorati, G., Ceci, R., Rivezzi, G., Cirasino, L., Carideo, P., Black, D. M., Garzillo, C., and Giani, U. (2013). A general model of dioxin contamination in breast milk: results from a study on 94 women from the caserta and naples areas in italy. *Int J Environ Res Public Health*, 10(11):5953–5970.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512.

Robins, J. M. (1987). Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". *Computers & Mathematics with Applications*, 14(9):923–945.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121.

Rocco, G., Petitti, T., Martucci, N., Piccirillo, M. C., La Rocca, A., La Manna, C., De Luca, G., Morabito, A., Chirico, A., Franco, R., et al. (2016). Survival after surgical treatment of lung cancer arising in the population exposed to illegal dumping of toxic waste in the land of fires ('Terra dei Fuochi') of Southern Italy. *Anticancer research*, 36(5):2119–2124.

Rosenbaum, P. R. (2002). *Observational Studies*, pages 1–17. Springer New York, New York, NY.

Roth, J. (2022). Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 4(3):305–22.

Roth, J., Sant'Anna, P. H., Bilinski, A., and Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Sandall, J., Tribe, R. M., Avery, L., Mola, G., Visser, G. H., Homer, C. S., Gibbons, D., Kelly, N. M., Kennedy, H. P., Kidanto, H., et al. (2018). Short-term and long-term effects of caesarean section on the health of women and children. *The Lancet*, 392(10155):1349–1357.

Sant'Anna, P. H. and Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122.

Santías, F. R., Cadarso-Suárez, C., and Rodríguez-Álvarez, M. X. (2011). Estimating hospital production functions through flexible regression models. *Mathematical and Computer Modelling*, 54(7-8):1760–1764.

Saviano, R. (2006). *Gomorra: Viaggio nell'impero economico e nel sogno di dominio della camorra*. Mondadori.

Schiele, F., Gale, C. P., Bonnefoy, E., Capuano, F., Claeys, M. J., Danchin, N., Fox, K. A., Huber, K., Iakobishvili, Z., Lettino, M., et al. (2017). Quality indicators for acute myocardial infarction: A position paper of the acute cardiovascular care association. *European Heart Journal: Acute Cardiovascular Care*, 6(1):34–59.

Selvaraj, R. C., Cioffi, G., Waite, K. A., Jackson, S. S., and Barnholtz-Sloan, J. S. (2025). A pan-cancer analysis of age and sex differences in cancer incidence and survival in the united states, 2001-2020. *Cancers*, 17(3).

Senior, K. and Mazza, A. (2004). Italian "Triangle of death" linked to waste crisis. *The Lancet Oncology*, 5(9):525–527.

SiVeAS (2014). L'erogazione dei LEA nelle Regioni in Piano di rientro: Trend 2007 - 2012 (in Italian). Technical report.

Siverskog, J. and Henriksson, M. (2022). The health cost of reducing hospital bed capacity. *Social Science & Medicine*, 313:115399.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231.

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

Svimez (2022). Promozione di iniziative di studio e ricerca socioeconomica a supporto della programmazione e valutazione delle politiche di coesione della regione Campania – Dinamiche demografiche e percorsi di sviluppo territoriale.

Torbica, A. and Fattore, G. (2005). The "Essential Levels of Care" in Italy: when being explicit serves the devolution of powers. *The European Journal of Health Economics*, 6:46–52.

Tortora, F. (1989). Affare rifiuti, il pci spiega come sottrarlo alla camorra. *Il Giornale di Napoli*.

Trinca, S., Comba, P., Felli, A., Forte, T., Musmeci, L., and Piccardi, A. (2001). Childhood mortality in an area of southern italy with numerous dumping grounds: application of gis and preliminary findings. In *European Conference Geographic Information Sciences in Public Health: 2001; Sheffield*, volume 19.

Vazquez-Bare, G. (2023). Identification and estimation of spillover effects in randomized experiments. *Journal of Econometrics*, 237(1):105237.

Webb, M. D. (2013). Reworking wild bootstrap based inference for clustered errors. Technical report, Queen's Economics Department Working Paper.

WHO (2006). Quality of care: a process for making strategic choices in health systems.

WHO (2018). *Handbook for national quality policy and strategy: a practical approach for developing policy and strategy to improve quality of care*. World Health Organization.

WHO (2024). *Chapter 4. Solid waste. In: Compendium of WHO and other UN guidance in health and environment, 2024 update*. World Health Organization.

Williamson, S., Landeiro, F., McConnell, T., Fulford-Smith, L., Javaid, M. K., Judge, A., and Leal, J. (2017). Costs of fragility hip fractures globally: a systematic review and meta-regression analysis. *Osteoporosis international*, 28:2791–2800.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141(2):1281–1301.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Wooldridge, J. M. (2021). Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators. *Available at SSRN 3906345*.

Xu, R. (2025). Difference-in-differences with interference.

Estratto per riassunto della tesi di dottorato

Studente: Ciaccio Andrea matricola: 956559

Dottorato: Economics

Ciclo: 36

Titolo della tesi :Essays in Causal Inference and Health Economics

Abstract: Il presente lavoro di tesi è finalizzato a evidenziare l'importanza dell'utilizzo di tecniche di microeconometria di frontiera nell'area dell'inferenza causale, con heterogeneous treatment effects, per rispondere a domande rilevanti da un punto di vista di policy. Nel Capitolo 1 analizzo l'impatto di una misura di contenimento dei costi, introdotta per la prima volta in Italia nel 2007, sulla qualità e l'efficienza dei Servizi Sanitari Regionali (RHS). Per tenere conto della natura staggered della politica, sfrutto il Two-way Mundlak device. Nel complesso, la politica ha ridotto i costi senza implicare un aumento dell'efficienza dei RHS. Nel Capitolo 2 fornisco un toolkit pratico per stimare la distribuzione dell'untreated potential outcome per il gruppo trattato, in contesti non sperimentali con adozione staggered del trattamento, generalizzando lo stimatore del quantile treatment effect on the treated proposto da Callaway e Li (2019). Nonostante sia leggermente distorto, per le dimensioni del campione relativamente piccole, le prestazioni del metodo proposto aumentano sostanzialmente quando aumenta la dimensione campionaria. Infine, il Capitolo 3 studia l'effetto dei crimini ambientali sugli esiti sanitari. Per farlo, stimo l'effetto causale della Terra dei Fuochi - un'area di 90 comuni tra le province di Napoli e Caserta, dove, a partire dalla fine degli anni '80, si sono verificati scarichi e incendi illegali di rifiuti (tossici) ad opera della Camorra - sugli esiti sanitari. Utilizzando uno stimatore doubly-robust, sono riuscito a rilevare eccessi di mortalità per cancro nella Terra dei Fuochi rispetto al resto della regione.

In this thesis, I show the importance of using frontier microeconometrics techniques in the area of causal inference with heterogeneous treatment effects to answer policy-relevant questions. Specifically, in Chapter 1, I investigate the impact of a cost-containment measure first introduced in Italy in 2007 on Regional Health Services (RHSs) quality and efficiency. To account for the staggered nature of the policy, I exploit the Two-way Mundlak device. Overall, the policy reduced costs without implying a boost in the efficiency of RHSs. In Chapter 2, I provide a practical toolkit to recover the distribution of the untreated potential outcome for the treated group in non-experimental settings with staggered treatment adoption by generalizing the existing quantile treatment effects on the treated estimator proposed by Callaway and Li (2019). Despite being slightly biased for relatively small sample sizes, the proposed method's performance increases substantially when the sample size increases. Lastly, Chapter 3 studies the effect of environmental crimes on health outcomes. To do so, I estimate the causal effect of the Land of Fires - an area of 90 municipalities between the provinces of Naples and Caserta, where, starting from the end of the 1980s, illegal dumping and burning of (toxic) waste has occurred because of Camorra - on health outcomes. Using a doubly robust estimand, I find an excess of cancer-specific mortality in the Land of Fires compared to the rest of the region

Firma dello studente

Andrea Ciaccio