Ca' Foscari University of Venice

Department of Economics

Working Paper

Andrea Ciaccio

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

ISSN: 1827-3580
No. 24/WP/2023

Working Papers
Department of Economics
Ca' Foscari University of Venice
No. 24/WP/2023
ISSN 1827-3580

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
Ca' Foscari University of Venice

Abstract
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 managed to exit, but seven are still treated – raising the question of whether cost reduction has had any unintended negative effect 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 models the staggered nature of the policy by allowing me to analyze how the treatment effect varies along different dimensions. Further, it allows 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 a set of bounded-variations assumptions.

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

JEL Codes
C14, C21, I10, I18, J38

Address for correspondence:
Andrea Ciaccio
Department of Economics
Ca' Foscari University of Venice
Cannaregio 873, Fondamenta S.Giobbe
30121 Venezia - Italy
e-mail: andrea.ciaccio@unive.it

The Working Paper Series
is available only on line
(http://www.unive.it/pag/16882/)
For editorial correspondence, please contact:
wp.dse@unive.it

Department of Economics
Ca' Foscari University of Venice
Cannaregio 873, Fondamenta San Giobbe
30121 Venice Italy
Fax: ++39 041 2349210

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*

Abstract

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 managed to exit, but seven are still treated – raising the question of whether cost reduction has had any unintended negative effect 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 models the staggered nature of the policy by allowing me to analyze how the treatment effect varies along different dimensions. Further, it allows 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 a set of *bounded-variations* 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, for their precious advice, insights, and detailed comments on previous drafts of the paper. I am also grateful to Domenico Depalo for the detailed comments received on a previous paper draft. I also thank the audience at the Ca' Foscari Internal Departmental Seminar, the 11^{th} SIdE Workshop for Ph.D. Students in Econometrics and Empirical Economics and the XXXV SIEP Conference for their helpful comments.

*Department of Economics, Ca' Foscari University of Venice, Italy. E-mail: andrea.ciaccio@unive.it

1 Introduction

The rising concern about the sustainability of public finances, and in particular, the need to respect budget balances, has aroused the attention of policymakers worldwide, especially after the onset of the 2007-2008 financial crisis. Some countries have implemented different measures to cut down public expenditure, particularly for healthcare, drastically. Processes aimed at rationalizing healthcare spending are, in fact, commonly undertaken to restore budget balance, especially in countries with publicly funded health services. The main reason is that countries usually devote much of their GDP to health expenditure. For instance, in OECD countries, before the COVID-19 pandemic outbreak, healthcare expenditure accounted, on average, for 8.8% of their GDP expenditure.

If, on one side, healthcare spending represents a natural candidate to curb public expenditure; on the other side, the increasing demand for access to healthcare facilities, coupled with the need for healthcare systems to provide high-quality care, has spurred numerous studies to investigate whether the financial crisis and, more specifically, budget cuts to healthcare systems may have adversely affected the health of the population (see, for instance, Quaglio et al. (2013), Franklin et al. (2017), Golinelli et al. (2017), and Arcà et al. (2020)), as higher health expenditure is usually associated with better health outcomes (OECD, 2021).

Thus far, the health community has not agreed upon a single and worldwide-accepted definition of quality care (WHO, 2006, 2018). One commonly employed definition, however, comes from 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). Moreover, the United States Institute of Medicine acknowledges that effectiveness, efficiency, equity, patient-centredness, safety, and promptness are crucial dimensions of care.

Besides providing high-quality services, healthcare systems need also to be efficient. Especially in countries with limited resources for healthcare, improving the efficiency of health systems is one of the major goals for policymakers to face the increasing demand for health services. It is, therefore, crucial to define, for this paper, what we mean by efficiency. Even though it is hard to measure efficiency in health – as what is crucial for patients are not outputs but outcomes – two broad categorizations of efficiency have been proposed over the years. The first is *technical efficiency*, which is concerned with maximizing the number of outputs/outcomes, keeping fixed the level of inputs. The second category, *allocative efficiency*, is concerned with how resources are allotted. More specifically, the aim is to achieve the best health outcome by incurring the least possible cost (OECD, 2016). For the rest of this paper, I will use the term efficiency to refer to allocative efficiency, as the purpose will be to understand whether budget cuts to healthcare have had any unintended consequences on health outcomes.

Italy represents an interesting scenario for understanding whether healthcare spending cuts may have unintended effects on health outcomes. Italy perfectly aligns with the average OECD country, spending around 8.7% of its GDP on healthcare (OECD, 2021). Moreover, despite the Italian National Health Service (NHS) being a multi-tier system characterized by substantial variability across regions – entitled to managing the Regional Health Service (RHS) – in terms of the quality of services provided (Aimone Gigio et al., 2018), the introduction at the beginning of the 2000s of Essential Levels of Assistance (ELAs) ensures comparability of healthcare provision across regions. In addition, although one of the main goals of having a decentralized system was to formally attribute regions the responsibility for

the RHS's financial stability, the Central Government has continued to finance ex-post the large deficit run by regional governments over the years. This has led public health expenditure and, in turn, the total deficit for public health expenditure to grow dramatically. In this context, *Piani di Rientro sanitari* (or Recovery Plans (RPs) in English) were first introduced in 2007 to counteract the increased spending and restore the financial stability of RHSs. The aim of this policy was twofold: on the one side, to restore budget balance, and on the other side, to guarantee (or improve) ELAs. Since these two targets are incoherent unless efficiency gains are feasible, as pointed out by Depalo (2019), it is legitimate to ask whether introducing such plans has had any negative consequences on health outcomes.

Sixteen years after the first *Piano di Rientro* (PdR) was signed, ten out of twenty-one RHSs have undergone at least one round of the policy. Three regions managed to exit, whereas seven remain under a Recovery Plan. There is a wide consensus that introducing such a policy reduced costs; by contrast, the effects in terms of quality and efficiency are mixed. On the one hand, treated regions report a deterioration in the quality of services provided after the plan's introduction (Calabrò, 2016). On the other hand, the Central Government registers positive effects regarding efficiency and no deterioration in the quality of the RHSs. However, some treated regions are still reported 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., Bordignon et al. (2020); Cirulli and Marini (2023)).

This paper aims to provide novel evidence regarding the causal impact of RPs on the quality and efficiency of RHSs, adding to the existing literature investigating the impact of budget cuts on healthcare on health-related outcomes (Heijink et al. (2013); Golinelli et al. (2017); Arcà et al. (2020) and many others). Specifically, it contributes to the existing literature in several ways. First, to the best of my knowledge, this is the first study evaluating the long-run impact of the policy. Despite two existing studies – Guccio et al. (2023) and Beraldo et al. (2023) – already employ data until 2018, as I do in the following analysis, they only provide an estimate of the overall effect of PdRs, thus failing to gauge the cumulative impact of the policy. Assessing the long-run effect is paramount to carrying out a sound policy evaluation. While on one side, effects on cost containment may be immediate to see – as regions undergoing a PdR are obliged to cut costs to access additional funding – this policy requires a structural reorganization of the RHS. As such, any sizeable effect on the quality of services provided and efficiency may take time to be detected, as also pointed out by Calabrò (2016). Besides, compared to Guccio et al. (2023) and Beraldo et al. (2023), I do not focus only on mortality rates or interregional patient mobility; this paper offers a more complete picture by also considering other indicators of quality, as well as efficiency and costs.

Second, the paper also contributes to the ongoing debate of whether *Piani di Rientro* managed or not to guarantee ELAs. Specifically, a wider set of quality indicators is considered. Besides considering some of the variables directly part of the MoH's ex-post monitoring, the effect of PdRs on other indicators is also evaluated. The policy might have indirectly affected these latter variables through reduced budgets available to treated regions. In particular, the hypothesis I want to test is whether regions have strategically outperformed on those indicators that discriminate on their ability to receive funds. If this is true, following the introduction of RPs – for treated regions – a deterioration of those indicators that are not part of the MoH evaluation process should be observed.

The third contribution of this paper will be to rely on the *DiD decomposition* proposed by Goodman-Bacon (2021), which allows the researcher to understand why the classical Two-Way Fixed-Effects (TWFE) estimator – which is commonly used in the existing literature of RPs – may not yield a con-

sistent estimator of the causal estimand of interest and what may be possible sources of bias. Different studies proved that the estimator for the average treatment effect on the treated (ATT) obtained by the TWFE is likely to be severely biased in a context with variation in treatment timing, as the one under scrutiny (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). Thus, failing to consider the staggered nature of the policy and the fact that the treatment effect is likely to vary over time and across regions may lead to invalid inference. Therefore, from an applied perspective, this study may also add to the strand of the literature on DiD (please refer to Roth et al. (2023) for a thorough review) by providing empirical evidence on the importance of incorporating the staggered entry set-up nature of the policy when carrying out program evaluation exercises.

Fourth, to avert the issue of having an inconsistent estimator of the ATT and to be able to estimate the causal effect of Recovery Plans, the staggered nature of the policy will be considered. In addition, heterogeneity of the treatment effect across regions and over time will be introduced. The former will allow us to understand whether regions that underwent a more abrupt reduction of costs experienced different impacts regarding the quality and efficiency of the RHS. The latter, instead, will let us estimate the long-run impact of the policy. To this end, I will exploit the estimator proposed by Wooldridge (2021), the *Two-Way Mundlak* (TWM). Although different studies have thus far evaluated the impact of Recovery Plans, to the best of my knowledge, this is the first paper that explicitly models the staggered nature of the policy and allows for heterogeneity of the treatment effect. The only two papers that, to some extent, introduced heterogeneity are Depalo (2019) and Beraldo et al. (2023). As far as the first paper is concerned, the author only considered regions first treated in 2007 and only evaluated the impact of the policy at the end of the first and second rounds of PdRs. Thus, being unable to consider that some regions left the treatment status over time and, simultaneously, to estimate the long-run effect. On the other side, Beraldo et al. (2023) exploited the estimator proposed by Callaway and Sant'Anna (2021) as a robustness check to understand whether the results obtained via their main identification strategy were in line with those obtained through an estimator that is robust to the presence of variation in treatment timing. Using Callaway and Sant'Anna (2021), the authors also estimate heterogeneous treatment effects; however, heterogeneity is only allowed over time and not across regions. Moreover, compared to the studies by Beraldo et al. (2023), I do not focus only on the escape rate but also consider other indicators. Lastly, I explicitly tackle the issue of having regions entering and exiting the treatment status at different points rather than assuming that it is irreversible.

Further, 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 employed.

Lastly, since three strong assumptions must be satisfied for the estimator of the ATT obtained via the TWM approach to be consistent, building on Depalo (2019), results obtained via the estimator proposed by Manski and Pepper (2018) will also be presented. This latter estimator will allow us to understand whether the results obtained using the parametric estimator are robust to the relaxation of 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. There are two main advantages of this approach. First, the exact reason the identifying assumption may not hold must not be known to the researcher. Second, it allows us to estimate region-specific treatment effects, thus shedding light on the results obtained via the parametric estimator.

I view the fact of performing inference using both the TWM and milder non-parametric assumptions as one of the strengths of this paper, as this allows me to understand how the results are sensitive to the relaxation of the identifying assumptions.

Overall, the main findings show that RPs managed to reduce costs. Cost reduction was pursued through decreased hospital beds and hospitalization rates. However, contrary to what was expected from the policymaker, except for Abruzzo and Calabria, the policy did not enhance the efficiency of treated RHSs. In addition, a deterioration in the quality of services was observed in regions that underwent a PdR. These results also hold in the long run and are robust to different sensitivity exercises. Moreover, the larger negative effects (in terms of quality) were documented in regions that experienced a more drastic reduction in the hospitalization rate.

This paper is structured as follows. Section 2 outlines the main features that characterize the Italian NHS and the process that led the Central Government to adopt Recovery Plans. In Section 3, how a Piano di Rientro works and the objective of the Central Government are exposed. Section 4 reviews the existing studies on RPs and, more in general, the literature in health economics analyzing the impact of cost-containment measures on health outcomes. In Section 5, I describe how the dataset used for the analysis was constructed and the indicators employed. Section 6 reports the methods used to estimate the causal effect of RPs, whereas Section Section 7 presents different sensitivity exercises. In Section 8, the main results obtained using both the parametric and the non-parametric approach are presented, as well as results obtained via different robustness checks. In this latter Section, the results obtained are also discussed in light of the existing literature. Lastly, Section 9 concludes.

2 Institutional Setting

This section briefly reviews the main features that characterize the Italian National Health Service (NHS) and the process that led the Central Government to adopt Recovery Plans to curb the excessive budget deficit for public health spending run by regional governments.

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. To comply with the Italian Constitution (*art. 32*), the system rests on three main guiding principles: 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 (MoH) 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

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

process of budgetary decentralization to settle the national deficit to comply with 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.

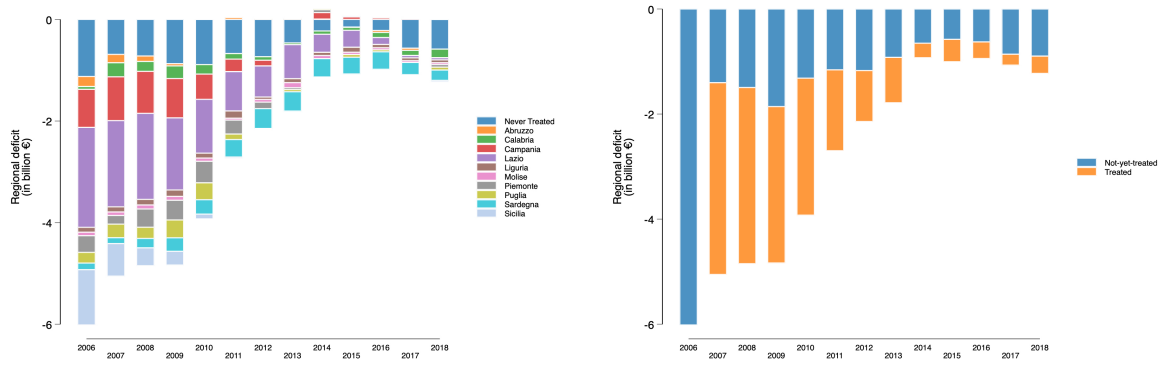
This decentralization process culminated with the *Riforma del Titolo V* in 2001, which formally attributed to regions the health protection and the responsibility to respect budget balance.[2] Moreover, to ensure each citizen was guaranteed a minimum level of health services and that these were comparable across regions, Essential Levels of Assistance (ELAs) 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 start growing dramatically over time.

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). In this context, the Central Government made effective, with the budget law for 2007, the *Piani di Rientro (PdRs) sanitari*.[5]

Figure 1: Regional deficit for public health expenditure



(a) (b)

Notes: The above 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.

[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.

[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.

3 Recovery Plans

Starting from 2007, regions running a deficit of over 5% (initially set at the 7%) of the overall level of funding were obliged to sign region-specific Recovery Plans to access financing.

An RP consists of a bilateral agreement between the region and the Central Government (in particular, the MoH and the Treasury), where the region must identify – via the analysis of the SWOT (strengths, weaknesses, opportunities, and threats) – all the necessary measures to be adopted to restore budget balance while guaranteeing (or improving) the ELAs, which are the RP's *general targets*. To reach general targets, in the RP are also included the *specific targets* in which the areas of interventions are outlined, and the *operational targets* that specify how the specific targets are pursued. Alongside the actions aimed at a structural reorganization of the RHS included in the plan, treated regions are likewise required to raise regional tax rates (*IRAP* and *IRPEF*) to increase revenues.

A PdR lasts three years; however, it may be renewed for an additional triennium (*round*) if the region does not achieve the predetermined goals. Moreover, in case of an enduring deficit or when objectives are significantly missed, a commissioner is automatically appointed.[6] In this latter case, more stringent restrictions regarding the RHS's tax rates and structural equipment are also enforced.

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 avoid a sudden drop in health-related spending.[7] For what concerns the support provided in managing and evaluating the RHSs, this is carried out by the MoH, in concert with the Ministry of Economics and Finance (MEF), through a monitoring system (the SiVeAS). This latter consists of an ex-ante monitoring – in which the Central Government inspects whether all the legal actions that need to be undertaken to restore financial stability are outlined in the plan – and an ex-post monitoring – through which the MoH verifies, every quarter, whether the budget balance is being restored and ELAs are being guaranteed. As also stressed by Depalo (2019), the centralization of this process is a salient aspect of this policy, as it ensures cross-regional comparability of the targets reached by each region.

The main targets of this policy are: on one side, to reduce management costs and, on the other side, to increase the efficiency of the RHS and the appropriateness of care provided. To reach these goals, regions signing a PdR were required to reduce the hospitalization rate and the number of hospital beds (per thousand inhabitants). Reducing the number of hospital beds was requested to facilitate a reconversion process towards more appropriate and, simultaneously, less expensive care than hospital admission. Conversely, following a reduction in the hospitalization rate, an increase in the percentage of hospital beds effectively occupied and a switch from ordinary to day surgery was expected. Although similar actions were implemented throughout the country with the budget law for 2007, the reduction treated regions were required to implement was more prominent, as their starting structural equipment was higher. Moreover, with the budget law for 2010, for regions under PdRs that (partly) miss the plan's target, the automatic hiring stop and turnover freezing became mandatory.

From 2007, ten out of twenty regions have undergone at least one round of Recovery Plans. The first plan was signed by the region Lazio in February 2007. In the same year, Abruzzo, Campania, Liguria,

[6] In addition, to ensure the correct execution of the plan, the Central Government may also nominate two sub-commissioners.

[7] Please refer to Depalo (2019) for an overview of how these funds were made available.

Molise, Sardegna, and Sicilia were also forced to join a PdR. Instead, Calabria, Piemonte, and Puglia signed an RP in the following years (December 2009, July, and November 2010, respectively). Liguria and Piemonte met RPs' requirements in April 2010 and March 2017, respectively, whereas Sardegna left the plan invoking its special statute in December 2010. On the other side, starting from 2008, some regions were put under the administration of an external commissioner. These regions are Lazio (2008-2020), Abruzzo (2008-2016), and Campania (2009-2020). The only two regions which, to date, are still under a commissioner are Calabria and Molise (from 2010 and 2009, respectively). The complete history of RPs' adoption in Italy is summarized in Table 1, in which the situation on September 1^{st} of a given year is depicted.[8]

As shown in Figure 1, Recovery Plans' aim at curbing excessive health-related spending was reached. The total deficit decreased from 6 billion euros in 2006 to less than 2 billion euros in 2018. However, it is unclear whether cost containment was pursued through an increase in efficiency or due to a deterioration of the quality of the services provided by RHS. Although from the second round of RPs,[9] the law gave both ELAs and financial stability the same importance, the two objectives are incoherent unless efficiency gains are feasible, as pointed out by Depalo (2019).

More specifically, despite the health community has not agreed upon a worldwide-accepted definition of quality care (WHO, 2006, 2018), one commonly-employed definition comes from 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). For the remainder of the paper, I will use "quality" to refer to any health service that may lead to better health outcomes.

If, on one side, there is no shared definition of quality in the health literature, on the other side, the concept of efficiency is even harder to define. According to the economic theory, efficiency is usually referred to as the relation between inputs (e.g., hospital supply) and either intermediate (e.g., waiting time) or final outputs (number of life saved, quality-adjusted life years) (Palmer and Torgerson, 1999). What is crucial for patients are not outputs but outcomes. For this reason, two broad categorizations of efficiency have been proposed over the years. The first is *technical efficiency*, which is concerned with maximizing the number of outputs/outcomes, keeping fixed the level of inputs. The second category, *allocative efficiency*, is concerned with how resources are allotted. More specifically, the aim is to achieve the best health outcome by incurring the least possible cost (OECD, 2016). For the rest of this paper, I will use the term efficiency to refer to allocative efficiency, as the purpose will be to understand whether budget cuts to healthcare have had any unintended consequences on health outcomes.

4 Related Literature

Different studies have tried to analyze the impact of RPs on cost containment and the quality of RHSs. The consensus is that implementing such plans effectively reduced costs (Aimone Gigio et al., 2018; Depalo, 2019; Arcà et al., 2020; Bordignon et al., 2020). Conversely, there are mixed findings regarding the impact of such a policy on the efficiency and quality of the services provided. This is mainly due to

[8] Results, however, are robust if the situation on June 1^{st} is considered. These results are available upon request.

[9] In the first edition of PdRs, the "implementation monitoring" carried out by the Central Government was mainly targeted to verify the region's budget balance.

Table 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: For each year, the situation on September 1^{st} is reported. **RP** = if the region is under PdR; **RPC** = PdR + presence of commissioner.

different model specifications and identification strategies used by previous studies.[10]

Exploiting a Two-way Fixed Effects (TWFE) estimator coupled with an IV strategy – using as an instrument the percentage of citizens complaining about the presence of waste in municipal streets – to clean for potential violation of the parallel trend assumption, Bordignon et al. (2020) find no statistically significant effect of PdRs on patients' health outcomes. In particular, the authors employ 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. On the other side, Cirulli and Marini (2023) – also relying on a TWFE estimator and using the percentage of citizens disappointed about public transport as an instrument – document an increase in the total mortality rate, suicide rate, mortality rate associated with cancer and heart diseases, and discharge rates due to psychological issues in treated regions following the introduction of RPs.[11] Similarly, exploiting a Matching Method for Time-series Cross-Sectional Panel coupled with a TWFE estimator, Guccio et al. (2023) estimate a negative impact of the policy on the total mortality rate. Beraldo et al. (2023), on the other side, using a TWFE estimator, find that introducing PdRs has had unintended negative consequences on the escape rate as well. Lastly, using a non-parametric approach, Depalo (2019) estimates a sudden drop in the total hospitalization rate – in treated RHSs – followed by an increase in the mortality rate (all causes) by the end of the first round of RPs. Specifically, the reduction in the hospitalization rate is documented to be larger in those regions experiencing a more considerable decline in health spending – besides, the larger the drop in hospitalization rate, the higher the increase in mortality.

For what concerns efficiency, instead, Depalo (2019) finds that RHSs' efficiency – proxied 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 PdRs. Similarly, resorting to a Data Envelopment Analysis (DEA), Giancotti et al. (2020) document that RPs did not impact hospital efficiency – as proxied by the technical efficiency index – in 2010-2013. In particular, this index is constructed as an output-to-input ratio, where hospital bed capacity and hospital staff are employed as inputs. The total number of inpatient discharges gives the output. Conversely, using a similar strategy to Giancotti et al. (2020) and data for 2003-2010, Guccio

[10] It is worth noting that all existing studies on RPs employ publicly-available data – obtained through the Health For All database or the MoH – which are already aggregated at the regional level. So, the unit of observation is RHS r at time t.

[11] It is worth stressing that the two sample periods are approximately the same. Whereas Bordignon et al. (2020) use data from 2000-2014, Cirulli and Marini (2023) employ only two additional years: 1999 and 2015.

et al. (2022) discover an increase in inefficiency – where cost inefficiency is estimated using Stochastic Frontier Analysis – in treated regions in the years following the introduction of such a policy.

All the studies mentioned so far investigating the impact of RPs on the quality of RHSs belong to a broader literature in health economics that analyzes the effect of healthcare spending cuts on efficiency and, in particular, on the quality of healthcare providers. For instance, using a 15-year panel on Italian RHSs, Golinelli et al. (2017) investigate the relationship between per-capita healthcare expenditure and the total mortality rate. Overall, they find that reducing healthcare spending is associated with an increase in the mortality rate, all else being equal. Similarly, using a 10-year (2004-2014) panel on Italian regions and exploiting the introduction of PdRs as an instrument to clean for potential endogeneity issues in healthcare spending, Arcà et al. (2020) estimate a 4.5% increase in avoidable mortality following a €100 cut in per-capita health expenditure. Such an increment in avoidable deaths seems to have been caused by a drastic reduction in hospital supply (beds and staff) and decreased hospitalization rates. Furthermore, the authors also document that patients' mobility flows – mainly from Southern regions to the Center-North of Italy – increase significantly following a decline in healthcare spending. This is probably due to a reduced hospital supply in Southern regions after introducing PdRs.

Inter-regional patients' mobility is, in fact, strongly affected by hospital supply. Using a dynamic spatial panel data model and Italian hospital discharge records for 2001-2010, Balia et al. (2018) discover that, ceteris paribus, a larger number of hospital beds in the region of origin decreases the likelihood of seeking care in another region. At the same time, the outflow toward a given region increases with hospital supply. If, on one side, an excess of hospital beds is likely to indicate bad management, waste of resources, and, consequently, low quality of services provided. Conversely, larger hospital bed capacity may also suggest lower waiting lists.

Regarding the impact of efficiency on health outcomes, Martini et al. (2014) employ 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. At the same time, more efficient hospitals also exhibit lower readmission rates.

The results presented above are not specific to the Italian context only. Using a panel on 14 Western countries from 1996-2006, Heijink et al. (2013) evaluate the relationship between healthcare spending and avoidable mortality. The authors report that countries with an expenditure growth above the average have a larger decrease in avoidable deaths. Similarly, using 2005-2006 data on primary care trusts in the UK and relying on a Generalized Method of Moments estimation procedure with several instruments, Andrews et al. (2017) document that mortality (all causes) is highly sensitive to variations in health spending. In particular, they estimate an elasticity of -0.705 of the mortality rate to variation in healthcare expenditure.

Analogously to what has been documented for Italy, employing 1996-2015 data from the Spanish Ministry of Health and Ruhm's Fixed Effects model, Borra et al. (2020) also report a significant increase in the mortality rate from circulatory diseases and external causes following a reduction in hospital supply. Similar results are found in Sweden by Siverskog and Henriksson (2022), who use 2001-2019 data for Swedish regions and a TWFE model to analyze the impact on health outcomes of reducing bed capacity. Despite bed capacity and deaths having declined nationwide in the period considered, the largest reduction in mortality rates is documented in regions exhibiting a less pronounced reduction in hospital beds. Furthermore, the authors also report that providing an additional bed generates

three additional quality-adjusted life years (QALYs). This latter finding corroborates the hypothesis that bed capacity is directly linked with patients' outcomes. Conversely, Cook et al. (2012), exploiting the introduction of the California Assembly Bill 394 – that mandated the maximum levels of patients hospital staff should nurse –, discovered that decreasing the patient-nurse ratio does not improve patients' health outcomes.

Overall, the existing studies suggest that measures aimed at reducing costs through reductions in healthcare spending and hospital supply may adversely affect patients' health. Moreover, these policies have usually proven ineffective in promoting health providers' efficiency. For these reasons, it is crucial to understand whether the introduction of PdRs has had any unintended effects on the quality of RHSs. To address this issue, a novel estimation procedure taking into consideration the complexity of the policy under study will be used (see Section 6). Furthermore, indicators that are both directly part of the ex-post monitoring and other indicators that such a policy may indirectly impact will be employed.

5 Data

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.[12] In contrast, all the other indicators of RHSs' quality and costs, as well as controls, were downloaded through the *Health for All (HFA)* software.[13] 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.[14] 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.

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

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

[14] 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.

5.2 Dependent Variables

Different proxies for quality, which may directly or indirectly affect the population's health conditions, are used to understand whether adopting an RP has had any consequences on the quality of RHS. Besides considering some of the variables directly part of the MoH's ex-post monitoring, the effect of PdRs on other indicators is also assessed. The policy might have indirectly impacted these latter variables through a reduced budget available to the region. In particular, the hypothesis I want to test is whether regions have strategically outperformed on those indicators which discriminate on their ability to receive funds. If this is true, following the introduction of RPs – for treated regions – a deterioration of those indicators that are not part of the MoH evaluation process should be observed.

The first proxy for RHSs' 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 3). They are commonly employed in the health literature as a proxy for healthcare utilization (Depalo, 2019; Arcà et al., 2020) and as an (indirect) indicator of Quality of Life in patients (Berchialla et al., 2010).

Next, since numerous studies document a negative (indirect) impact of healthcare spending cuts on deaths (Heijink et al., 2013; Andrews et al., 2017; Golinelli et al., 2017; Arcà et al., 2020), the *(total) mortality rate* is also considered as a proxy for RHSs' quality. The mortality rate under consideration is measured as the number of deaths (all causes) in a given year divided by the region's population size multiplied by $10,000$.

Another indicator of quality that is used in the following analysis is the *mortality rate from ischaemic heart diseases* – defined precisely as the mortality rate reported above.[15] Despite this indicator is not directly part of the ex-post monitoring carried out by the Central Government, different studies employ the Acute Myocardial Infarction (AMI) mortality rate – which is incorporated in the indicator under scrutiny – as an indicator of the quality of the healthcare provider (Pross et al., 2017; Schiele et al., 2017). Besides, although mortality due to coronary heart disease has substantially decreased over the last decades, AMI is still among the major causes of death in many OECD countries, including Italy (OECD, 2021). For this reason, it is crucial to understand whether PdRs had any impact on such an indicator. In particular, whether the decrease in the mortality rate from ischaemic heart diseases slowed down following the introduction of this policy.

The free-patient choice is a salient feature of the Italian System, where patients can seek care in hospitals outside the home RHS.[16] Significant inter-regional patient mobility in the ordinary regime characterizes the Italian NHS – equal to 7.2% of total hospitalizations only in 2020 (MoH, 2020) – with the Central and Northern regions of Italy being net exporters of hospital care (Balia et al., 2018). The large fraction of patients from the South seeking care in the rest of Italy indicates enduring differences in RHSs' quality. Since "the compensation of net patient flows has generated additional amounts of financial resources in favor of central-northern regions, and has exacerbated the north-south gradient in the Italian NHS" (Balia et al., 2018, p.3), it is interesting to analyze 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*

[15] In particular, this indicator consists of the following diseases: 410-414 (according to the ICD-9-CM classification); and 120-125 (according to the ICD-10-CM classification).

[16] Despite patients being required to bear the cost of traveling to another region, the residence region will pay the reimbursement for the treatment received according to a compensation scheme based on DRG tariffs.

within 48 hours in an ordinary regime (in short, *% of patients with hip fracture replacement*) is also estimated. Numerous studies in the health literature employ this variable as an indicator of the healthcare provider's quality (Pross et al., 2017; OECD, 2021), as the quality of life declines significantly after hip fracture (for instance, see Amarilla-Donoso et al. (2020)). Besides, the % of patients with hip fracture replacement is commonly used as a proxy for the healthcare system's responsiveness. Evidence suggests that promptly intervening with surgical procedures – within 48 hours from the accident – enhances patient health and reduces the risks of complications (OECD, 2021). This is because the risk of death and pressure sores are found to increase with surgical delay remarkably (Moja et al., 2012).

To test whether RPs had any effects on RHSs' 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*) is employed as a proxy for the inefficient use of healthcare services. Despite it is recognized in the health literature that it is hard to measure healthcare providers' efficiency (McGuire, 1987), inappropriate hospitalizations are commonly used to indicate the misuse of resources (Angelillo et al., 2000; Navarro et al., 2001; Pileggi et al., 2004). Moreover, the MoH acknowledges this variable as an indicator of the appropriateness and efficiency of services provided; it represents one of the indicators constituting the *Monitoraggio Griglia Lea* reports.[17]

Another valid proxy for the inappropriate use of resources, commonly used 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). There are two motivations for this indicator to be used as a proxy for the inappropriateness of care. First of all, although both the two delivery modes are not risk-free (Gregory et al., 2012), c-sections are found to be associated with increased risk of mortality for both the mother and the baby, increased risk of complications, as well as negative short- (e.g., increased likelihood of developing allergy and asthma) and long-term effects (child cognitive development) for the baby. This is because c-sections are a more invasive procedure (Polidano et al., 2017; Sandall et al., 2018; OECD, 2019). Moreover, c-sections represent a more costly procedure than vaginal birth.

The MoH also acknowledges the c-section for first-time mothers as a valid proxy of the appropriateness and efficiency of services provided. It is part of the indicators constituting the *Monitoraggio Griglia Lea* reports. Unfortunately, since this latter indicator is available only for 2012-2016, the *% c-sections* – defined as the number of c-sections over the total number of childbirth multiplied by 100 – will be employed as a proxy.[18]

For what concerns costs, two indicators that are both part of the ex-post monitoring carried out by the Central Government are employed, the *log* of *current health expenditure* and the *number of hospital beds*. The former is directly linked to the healthcare expenditure run by the regional government and the definition of budget balance imposed by the Maastricht Treaty's rules (Eurostat, 2013). The latter represents another channel through which management costs can be directly reduced (Aimone Gigio et al., 2018; Arcà et al., 2020).

[17] The complete list of DRG classified at high risk of inappropriateness is made available by the Italian MoH. Some examples of DRGs considered at risk of inappropriateness can be found in Appendix A.

[18] To check whether these two indicators convey the same information, I regressed the indicator provided by the MoH (% c-section for first-time mothers) on the one coming from the HFA (% c-sections), controlling for region and year fixed effects. The R^2 equals .90, suggesting that these two indicators are strongly correlated.

5.3 Controls

A set of socio-economic variables that might explain 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*. It is worth stressing that, as stated in Section 2, population size and its distribution by age and gender represent the criteria for which the Central Government allocates funds to the regional jurisdictions.

Following Depalo (2019) and Balia et al. (2018), *GDP per capita* is also controlled for, as it allows for accounting for income effects. At the micro-level, GDP might be a good proxy for patients' ability to seek care in the private health sector following a reduced quality of services provided in the public hospital. On the other hand, at the macro-level, GDP might capture that richer RHSs may provide more appropriate and higher-standards services.

5.4 Potential Effects & Channels

The main objective of RPs is to restore budget balance while guaranteeing the ELAs. To reach this goal, treated RHSs were required to decrease hospital bed capacity. The idea was that lowering the hospital supply might reduce the incidence of inappropriate hospitalizations, thus favoring a reconversion towards more suitable and less costly procedures than hospital admission. However, since budget cuts to healthcare can leave unaltered the quality of services provided only if the efficiency of the healthcare system improves, it is worth understanding through which channels RPs may have (in)directly impacted the variables discussed above and, in particular, the potential effects we may observe.

As far as the impact on the hospitalization rate is concerned, if, on one side, a reduction in hospitalization rate may suggest that PdRs were effective in cost-containment, as this is one channel through which this policy works (please refer to Section 3); conversely, a decrease in hospitalization is usually associated with deteriorated health outcomes (Berchialla et al., 2010; Depalo, 2019; Arcà et al., 2020).

If, on one side PdRs directly affect hospitalization rates, on the other side, the mortality rate from ischaemic heart diseases is likely to be indirectly impacted by the policy. There are two main channels through which it is expected that PdRs may have affected this latter indicator. First, the complexity of the healthcare management and the high risk of re-hospitalization make AMI, and more in general ischaemic heart diseases, particularly costly for healthcare systems (Kwok et al., 2018; Lobo et al., 2020). Therefore, it could be the case that healthcare spending cuts may have adverse (indirect) effects on these outcomes through reduced budgets available to the RHS. Second, previous analyses document that AMI is sensible to variations in hospital supply. Increasing hospital beds – one of the channels through which PdRs work – is associated with a lower 30-day readmission rate after AMI (Brown et al., 2014).

As far as the % of patients migrating to other regions for ordinary acute hospitalization, if, following the introduction of RPs, larger outflows are detected in treated regions (which are mainly regions from the South of Italy), this would indicate that the perceived quality in the South has deteriorated and that PdRs have worsened the gradient between the South and the North of Italy in terms of quality of the NHS. There are two possible channels through which RPs may indirectly affect patient mobility. First, spending cuts in healthcare expenditure increase the flows from South to North of Italy (Arcà et al., 2020). Second, there is evidence that the lower supply in the region of origin, the higher the likelihood of seeking care in another region (Balia et al., 2018).

On the other hand, it is unclear whether PdRs may have adversely affected the percentage of pa-

tients with hip fracture replacement. On one side, like AMI, also hip fracture repair constitutes a significant burden for the healthcare system (Williamson et al., 2017). Therefore, healthcare spending cuts are likely to indirectly affect this indicator through reduced budget made available to the hospital and through lower hospital supply (beds and staff). The timely response from the healthcare provider is found to be influenced by hospitals' operating theatre capacity, flow, and access (OECD, 2021). Conversely, the % of patients with hip fracture replacement is one of the variables part of the *specific targets* outlined in RPs. Thus, regions have incentives to monitor this indicator continuously and to ensure adequate levels of this variable are reached to access financing.

Similarly, whether PdRs may have increased or decreased the cesarean rate is unclear a priori. On the one side, 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 for this indicator are reached. In addition, c-sections represent a more costly procedure than vaginal birth. Therefore, it could be that a reduction in the % c-sections after the introduction of RPs is detected. On the other side, the introduction of PdRs may have positively affected the % c-sections. One reason to expect a positive impact of the policy on such indicator is that, in Italy, cesarean rates are larger in private hospitals than in autonomous public ones (De Luca et al., 2021). If an increase in the % c-sections is documented following the introduction of the policy, then this could indicate that the perceived quality of healthcare services provided by public hospitals has deteriorated, pushing more women to seek care in private ones.

Lastly, hospital beds are usually a proxy for the capital factor used in hospital production (Santías et al., 2011; Giancotti et al., 2020; Guccio et al., 2022). Therefore, on the one hand, if reductions in the number of beds are detected (in treated regions) after introducing such a policy, this would indicate that the policy was effective in reducing costs. On the other hand, reduced beds might also imply a lower capital factor, which may, in turn, negatively affect patients' health conditions (Arcà et al., 2020; Borra et al., 2020; Siverskog and Henriksson, 2022).

6 Empirical Strategy

6.1 DiD Decomposition & Two-Way Mundlak Approach

As pointed out in Section 3, RPs were first introduced in 2007. While seven regions were forced to sign a PdR in 2007, the remaining three treated regions underwent a PdR in the years following 2007.[19] This framework is often called a *staggered treatment adoption* set up 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 average treatment effect on the treated (ATT) in periods where the treated regions are effectively under Recovery Plans.

Most existing studies on Recovery Plans rely on some variation of the classic Two-Way Fixed Effects (TWFE) estimator – such as the one presented in the following equation – to identify the impact of RPs on health-related outcomes and interpret the estimated coefficient for the treatment dummy as an

[19]For the sake of simplicity, in the following sections, the words Regional Health Service and region are employed indistinguishably to refer to an RHS. However, one should always remember that there exist 20 regions in Italy, whereas the number of RHSs is 19, plus the 2 autonomous provinces of Bozen and Trento.

estimator of the ATT. For the remainder of this paper, Wooldridge (2021)'s notation will be borrowed. Let us consider the following regression model:

$$y_{g,t} = \beta w_{g,t} + c_g + \eta_t + u_{g,t} \tag{1}$$

where $y_{g,t}$ is (one of) the outcome(s) of interest (discussed in Section 5) for region g at time t, with $g = 1, \ldots, 21$, and $t = y^{first}, \ldots, 2018$, and $y^{first} = min\{t : y_{g,t} \neq \text{NA} \quad \forall g\}$. That is, y^{first} represents the first year in which the dependent variable of interest is observed (non-missing) for all the RHSs. Let the random variable $w_{g,t}$ denote a binary treatment with support in $\{0, 1\}$. Since the main goal of this paper is to evaluate the impact of Recovery Plans, $w_{g,t} = 1$ will then denote whether region g at time t was under an RP.[20] Instead, c_g represent unit-specific fixed effects (FE), η_t are year FE, and $u_{g,t}$ is the error term.

Unless one is willing to make strong – and often implausible – assumptions that go beyond a generalization of the traditional parallel trends assumption to the setting with variation in treatment timing, it can be proved that the estimated coefficient for $w_{g,t}$ obtained via TWFE is an inconsistent estimator for the causal estimand of interest in a context with staggered policy rollout (Borusyak et al., 2021; Goodman-Bacon, 2021; De Chaisemartin and d'Haultfoeuille, 2022a). 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 of the ATT in the context with variation in treatment timing, please refer to Appendix B.

To estimate the causal effect of *Piani di Rientro* on the quality and the efficiency of services provided, I will resort to the *Two-Way Mundlak Approach* (TWM) first proposed by Wooldridge (2021). Specifically, Wooldridge, by proving the equivalence between the TWFE and the Pooled OLS (POLS) estimator applied to a regression that includes time-specific cross-sectional averages and unit-specific time averages, shows that it is still possible to retrieve a consistent estimator of the ATT within the simple regression framework – also in a context with variation in treatment timing – by simply introducing heterogeneity. Indeed, similarly to other recently developed estimators (e.g., Callaway and Sant'Anna (2021)), this approach allows the researcher to estimate heterogeneous treatment effects, where heterogeneity can be across treatment cohorts, intensities, calendar time, and/or covariates (see Appendix B.3).

Although different studies have thus far evaluated the impact of Recovery Plans, to the best of my knowledge, none of the existing studies – except Depalo (2019) to a certain extent – have allowed heterogeneity. From the contextual perspective, there are two main reasons why allowing heterogeneity across regions and over time may be interesting. First, for what concerns heterogeneity over time, while the salient features of the policy have remained fairly stable over time, the law attributed the same importance to ELAs and financial stability only from the second round of RPs (the first round of PdRs mainly focused on restoring the budget balance of RHSs'). In addition, while on one side, effects on cost containment may be immediate to see – as regions undergoing a PdR are obliged to cut costs to access additional funding – this policy requires a structural reorganization of the RHS. As such, any sizeable effect on the quality of services provided and efficiency may take time to be detected, as

[20] Please note that the reason for using g – rather than i – to refer to a specific region is twofold. On the one hand, this avoids confusion regarding the level at which the variables were computed since they are already aggregated at the RHS level. On the other hand, 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)).

also pointed out by Calabrò (2016). Second, while there is no specific reason to expect different effects across treated cohorts, PdRs are region-specific agreements. Therefore, it may be reasonable that, for instance, the effect for Calabria is different from the one for Lazio. For this reason, it may be worth allowing heterogeneity across regions.

To further corroborate the choice to rely on techniques that explicitly consider the staggered nature of the policy, the DiD decomposition proposed by Goodman-Bacon (2021) will be used (see Appendix B.2 for a review). Specifically, this procedure will allow us to understand why the TWFE estimator may not represent a consistent estimator of the ATT in a context with variation in treatment timing and what may be possible sources of bias.

Let us consider the following regression model:

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

using the same notation introduced above, $y_{g,t}$ represents (one of) the outcome(s) of interest, $w_{g,t}$ is a dummy variable taking value equal to 1 if region g at time t was under an RP, α represents the constant, η_t are year Fixed Effects (FE), whereas $d_{g,r}$'s are mutually-exclusive cohort dummies indicating when region g first received the treatment. Since new treated regions are entering the treatment only in 2007, 2010, and 2011, then $r \in \{2007, 2010, 2011\}$, meaning that $d_{2007} + d_{2010} + d_{2011} + d_{never} = 1$. Lastly, $intens_{g,t}(s) = \mathbb{1}\{t - e_g = s\}$ represents an indicator variable (i.e., a dummy variable) for region g at time t being s periods away from when it was first treated, and $e_g = min\{t : w_{g,t} = 1\}$.

Equation (2) can be estimated via POLS. It can be proved that τ_s in (2) is the estimator of the ATT (Wooldridge, 2021), where, in this specific case, the ATT is allowed to vary by treatment intensity. Allowing the treatment effect to vary by treatment intensity will enable the estimation of the long-run effects of RPs. This is paramount since, as explained above, the policy is structural, and therefore, it may take time for regions to adjust to PdR-specific requirements as well as for any sizeable effect to be detected.

However, for τ_s to be a consistent estimator of the ATT, three assumptions need to hold. The first assumption requires the treatment 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 second, instead, rules out anticipatory behaviors.

Before stating the following assumption, further notation should be introduced. Let q be the first period in which the policy is implemented; then, for $r \in \{q, \ldots, 2018\}$, $y_{g,t}(r)$ 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).

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

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

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.

Lastly, the third assumption needed for identification is a generalization of the parallel trends as-

sumption to the multi-period setup with staggered entry.

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

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

where $t = \{y^{first} + 1, \ldots, 2018\}$, and, as explained above, y^{first} is first period in which y is observed for all regions. 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, following Wooldridge (2021), the next regression will also be estimated:

$$\begin{aligned} y_{g,t} =& \alpha + \sum_r \lambda_r d_{g,r} + \sum_{s=0}^{11} \tau_s \left(w_{g,t} \cdot intens_{g,t}\left(s\right)\right) + \mathbf{x}_g \boldsymbol{\kappa} + \sum_r (d_{g,r} \cdot \mathbf{x}_g)\boldsymbol{\gamma}_r + \\ & (w_{g,t} \cdot \dot{\mathbf{x}}_{g,treat})\boldsymbol{\rho}_{treat} + \eta_t + u_{g,t} \end{aligned} \tag{3}$$

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] As explained in Section 5, $\mathbf{x}_g$ includes the region's population size and its distribution by age and gender and GDP per capita. 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.

Please note that differently from (2), in (3) the ATT's are allowed to vary, not only across treatment intensities but also with covariates. However, in this case, standard errors should be adjusted for the sampling variation in $\boldsymbol{\mu}_{treat}$.[22]

Introducing $\mathbf{x}$ and allowing the ATT to vary also with the covariates allows the researcher to relax both the PT and NA assumptions and to require them to be valid across subpopulations sharing the same characteristics.[23] That is:

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

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

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

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

As far as standard errors are concerned, it is common practice in the DiD literature to cluster standard errors at the level at which the treatment is assigned (if known) to account for the within-cluster correlation of the error term (Bertrand et al., 2004; Cameron and Miller, 2015; MacKinnon et al., 2023). As pointed out by Cameron and Miller (2015), indeed, including FE at the assigned treatment level is insufficient to purge for all the intra-cluster correlation of the disturbances. Since the treatment assign-

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

[22] To do so, the `margins`'s option, `vce(unconditional)`, is used. This option estimates standard errors allowing for the sampling variation of the x's.

[23] Please note that the TWM approach is similar to the estimator proposed by Sun and Abraham (2021). Still, compared to their estimator, it has one main advantage. As in Wooldridge (2021), the TE is explicitly allowed to vary with covariates, making the parallel trends assumption more credible.

ment mechanism is known in the setting under analysis, in regressions (2) and (3), standard errors will be clustered at the RHS level to correct for the potential serial correlation of the error term.

However, as explained in Section 3, in the setting under analysis – depending on the year considered – there are 10 treated RHSs at most. 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 overcome this issue, for each of the two regressions above, both classic cluster-robust standard errors, as well as standard errors based on the *subcluster wild bootstrap* with Rademacher weights (MacKinnon and Webb, 2018), will be computed. In this latter case, standard errors will be obtained using a version of the CRVE. Then the resampling procedure will be carried out at the region-year level (each pseudo-residual will contain 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 the subcluster wild bootstrap may lead to improved finite-sample inference in a context where the number of treated clusters is small. Please refer to Appendix B.4 for a review of how this procedure works.

6.2 Threats to Identification

For the estimator of the ATT obtained via the TWM to be consistent, the three (strong) assumptions presented in Section 6.1 need to hold. More specifically, besides requiring the treatment to be irreversible, the TWM hinges upon a generalization of the PT and NA assumptions – also known as *invariance assumptions* in the literature of causal inference (Appendix B.1) – to build the counterfactual outcome region g would have experienced at time t had a PdR not been signed by period r. The benchmark estimator works poorly if the PT or NA (conditional or unconditional) assumption fails. As a consequence, the estimator of the treatment effect is inconsistent, and all inference is unreliable.

Despite the great advantage of allowing point identification of the treatment effect, invariance assumptions – such as the PT and NA assumptions – are hard to justify in most empirical settings, as they require constructing the counterfactual with certainty.

In such a context, while it is true that assuming that the treatment is at an absorbing state is a strong requirement, as there are three switchers in such a context (Liguria, Piemonte, and Sardegna), the NA assumption is probably the more problematic to justify. If, on one side, it is easy to rule out the fact that regions first treated in 2007 could have hardly forecast the introduction of Recovery Plans, as all of the measures aimed at curbing the excessive spending run by local governments implemented in the years before PdRs had never become effective, as explained in Section 2, conversely, motivating the absence of anticipatory effects of later-treated units is harder. For instance, Calabria had already required the activation of a PdR in 2007, but this latter became effective only in December 2009. This delay was mainly because significant misalignments were present between the region's balance sheets and those sent to the MoH for PdR's activation. 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 check whether the NA is likely to be violated and, therefore, undermines the credibility of the results obtained through the TWM, different sensitivity exercises that tackle the issue of anticipatory behaviors will be run (see Section 7).

The PT assumption, on the other side, is one of the most debated assumptions in applied works. In

most settings, trends are likely already present before the policy implementation. Several procedures have been proposed to test whether such an assumption will likely hold empirically. For instance, Wooldridge proposes, within the POLS framework, a way to empirically test whether PT is likely to hold (i.e., to test for the presence of pre-trends). The test, which is based on exclusion restrictions, requires the introduction of intensity dummies in the pre-treatment period ($s < 0$) in equations (2) and (3). Failing to reject the null hypothesis of joint insignificance for the coefficients of the intensity dummies for $s < 0$ will indicate the absence of pre-trends.

However, tests like the one proposed by Wooldridge (2021) are pre-tests on identifying assumptions. There exists a wide literature showing that these tests feature several limitations. First, even if the PT is satisfied in the pre-treatment period, this does not guarantee that the post-treatment PT is also satisfied (Rambachan and Roth, 2023). Second, failing to reject the null hypothesis of the absence of pre-trends could be due to the low power of the test rather than the absence of such trends. This will invalidate inference, as any DiD-type estimator's consistency is based on valid PT. Lastly, conditioning the analysis on whether the test fails to reject or not the absence of pre-trends induces a selection bias, called the *pre-test bias* (Roth, 2022).

Besides the limitations of these tests, there has yet to be a consensus in the existing literature on what to do if pre-trends are detected. As such, the analysis should not be conditioned on test results only but rather motivate the choice with the researcher's context-specific information on why the PT may likely hold.

To check whether the PT is likely to be satisfied in this context, I depict in Figure 2, for each dependent variable of interest, the average evolution of each of the three treated cohorts against the never-treated. What emerges is that the PT is likely to be satisfied for only some of the dependent variables of interest (e.g., the mortality rate for ischaemic heart disease).

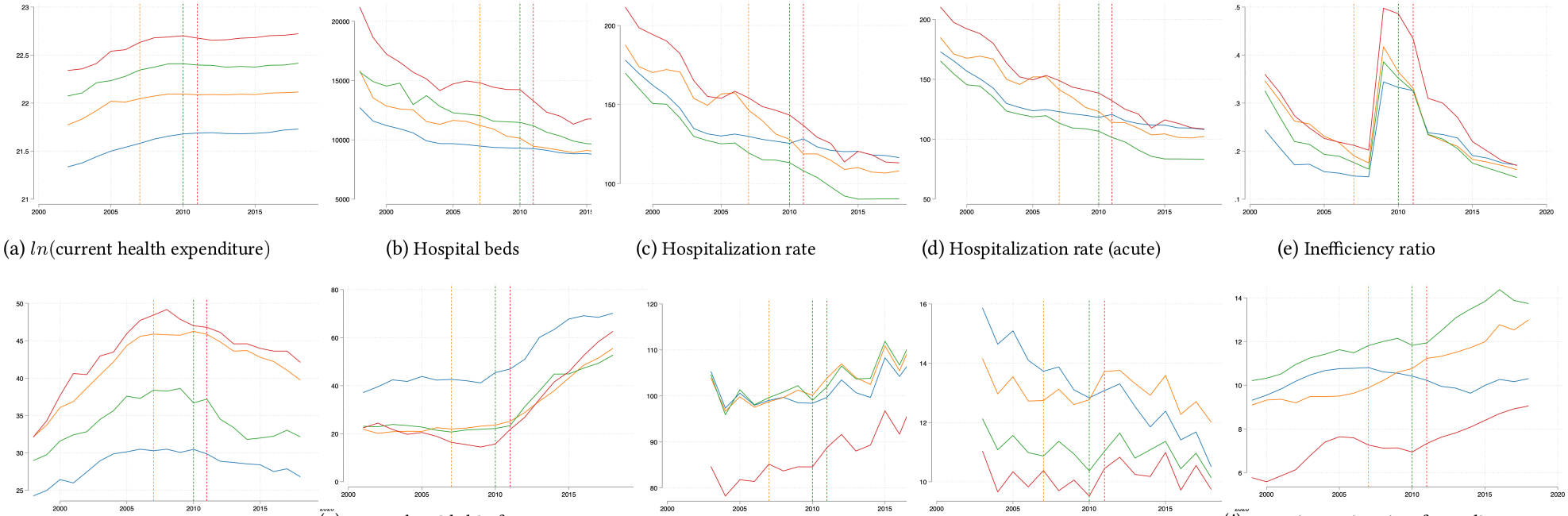
As it is clear from Figure 2, the PT is not likely to hold unconditionally in the context under analysis (at least for some of the outcomes of interest). One advantage of the TWM approach is that it allows relaxing the PT and NA assumption introduced above by relying on a set of covariates and allowing the treatment effect to vary with them. This is exactly what has been done in (3), where a full set of regressors, denoted by $\mathbf{x}$, have been introduced.

However, since only a few characteristics have been controlled for, these are likely insufficient to 'credibly' relax the PT and NA assumptions. If any of these two assumptions is severely violated, then the estimators of the ATT's are inconsistent.

One way to overcome the issue of either PT or NA assumptions being invalid is to rely on milder non-parametric assumptions, as proposed by Manski and Pepper (2013, 2018). The estimator they propose is, indeed, robust to the presence of anticipatory behaviors or the existence of pre-trends, as it will be explained in Subsection 6.3. More specifically, to enhance the inference's credibility, Manski and Pepper (2013, 2018) suggest giving up invariance assumptions in favor of assumptions of *bounded variation*, which bound the absolute difference between the true benchmark and the one built invoking an invariance assumption (Appendix B.1) not to exceed a given threshold, sometimes denoted with δ.

One last assumption, done implicitly, is that I am not distinguishing between regions under an external commissioner and those not. As explained in Section 3, from 2008, some regions were put under the administration of an external commissioner. However, since 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, not differentiating among regions under a commissioner or not should represent per se

Figure 2: parallel trends – 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 Panel (a) to (i) report the average evolution for each dependent variable of interest, split by treatment cohort. Specifically, the *blue* line represents the average evolution for the never-treated regions (i.e., regions that have never undergone a PdR over the observed period). The *orange* line is the average evolution for those regions that first signed an RP in 2007 (Lazio, Abruzzo, Campania, Liguria, Molise, Sardegna, and Sicilia). The mean evolution for regions entering an RP in 2010 (Calabria and Piemonte) is shown in *green*. The *red* line represents the evolution of those RHSs that first signed an RP in 2011 (Puglia). The *orange dashed* vertical line is drawn in correspondence with 2007, the *green* in correspondence with 2010, whereas the *red dashed* vertical line in correspondence with 2011 – which are the three years in which regions enter a PdR.

a threat to the identification. Also, this last issue can be easily tackled by resorting to bounds. This estimator allows the researcher to estimate region-specific treatment effects (TE), thus distinguishing between regions and those not under a commissioner in a given year.

Lastly, being able to estimate region-specific treatment effects allows me to overcome the issue that I needed to impose – when estimating (2) and (3) – that the treatment was at an absorbing state also for those three regions, which effectively left the treatment status over the sample period. Despite often "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), it may be too strong assuming that the region behaved as if it were still under treatment. If one thinks this assumption is too strong, one can completely ignore the estimated bounds for the TE for these three regions, as explained further below.

6.3 Bounds

To give the intuition behind the non-parametric approach proposed by Manski and Pepper (2018), I will follow Depalo (2019). Suppose the counterfactual outcome, $y_{g,t}(\infty)$, needs to be estimated. 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 $pre \in \{y^{first}, \dots, q\}$ 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 $\hat{y}_{g,t}(\infty) = \mathbb{E}(y_{never,t}(\infty))$. One can also invoke a "suitable" *parallel trends* assumption to estimate the benchmark as the ones shown in Section 6.1. 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 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 short overview of how to derive the counterfactual outcome and, in turn, the average treatment effect, please refer to Appendix B.1.

If any invariance assumptions hold, the benchmark outcome can be identified, and so the treatment effect. Conversely, suppose reasons exist for the appropriate exact invariance assumption not to hold. In these cases, Manski and Pepper (2013, 2018) suggest directly accounting for the uncertainty by requiring the invariance assumption to hold only approximately. That is, the benchmark against which $y_{g,t}(r)$ should be compared will be at most equal to the counterfactual estimated by relying 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, $\hat{y}^{Low}_{g,t}(\infty) = \hat{y}_{g,t}(\infty) - \delta$. By doing so, bounds for $y_{g,t}(\infty)$ are identified (Manski, 1990).[24]

There could be several reasons for exact invariance assumptions not to hold, such as pre-trends

[24] Please note that, in principle, there is no reason for δ to be symmetric. However, allowing $\delta^{Up} \neq \delta^{Low}$ "would constrain the spectrum of possible answers" (Depalo, 2019, p. 6) in terms of economic models one wishes to test. For this reason, in the following analysis, only symmetric δ's will be considered.

between treated and control units before the policy implementation, omitted variables, anticipatory behaviors by treated units, etc. However, one of the greatest advantages of the estimator proposed by Manski and Pepper (2018) is that the true reason why the invariance assumption may not hold should not be known for the estimator of the TE to be consistent. The only requirement is that the amount of uncertainty should be specified ex-ante. The larger the values of δ, the weaker the reliance on the invariance assumption, and the more credible the estimator derived (Manski, 2003).

Since the true values of the δ are unobservable, one way to choose its "optimal" value is to set δ equal to the (absolute) difference observed in the pre-treatment period between the actual outcome and the benchmark built invoking one of the assumptions. Manski and Pepper (2018), for instance, use the 75^{th} percentile of the absolute difference observed in the pre-treatment period. Similarly to Depalo (2019), in the following analysis, δ will be set equal to the largest 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.

Overall, by letting δ vary, a full set of possible counterfactual outcomes is identified. This implies that the treatment effect for region g that first underwent an RP in period r computed at time t will be bounded to lie in the interval $[\tau^{Low}_{g,r,t}, \tau^{Up}_{g,r,t}]$, where:

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

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

where $r \in \{2007, 2010, 2011\}$ 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 (SC) assumptions will be jointly exploited to estimate the causal effect of Piani di Rientro. In particular, the SC assumption is preferred over a state-invariance or a DiD assumption for one main reason. It allows constructing the benchmark employing all the information available (for untreated regions) in the pre-treatment, rather than relying only upon a simple average of never-treated regions' outcomes at period t (state invariance) or the last pre-treatment period (classic (2×2) DiD setup). Further, how SC constructs the counterfactual is similar to how the TWM does. However, differently from Depalo (2019), a longer time horizon will be considered, thus allowing the estimation of the long-run impacts of PdRs.

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:

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

[25] Please note that with assumptions of bounded variation, it is now possible to estimate the treatment effect for each treated region. This is surely one of the greatest advantages of relying on this approach. 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 $j \in \{Low, Up\}$.

For $y_{g,t}^{Low}(\infty) \leq y_{g,t}(\infty) \leq y_{g,t}^{Up}(\infty)$ 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 – since the lower bound could be greater than the upper bound – 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, the issue of how to perform inference in this scenario is not specifically addressed in the current work. This is mainly because the data employed are not a random sample from the population but the whole population itself. Moreover, Manski and Pepper (2018, p. 234) assert 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 overcome this issue would be to view current Italy as the realization of a sampling process. But this would require defining a super-population and a stochastic process generating the actual history of Italy. However, no consensus exists in the current literature on bounds on how to proceed in such cases.

For instance, placing restrictions on the possible post-treatment violation of the parallel trends assumption, Rambachan and Roth (2023) propose two approaches that allow obtaining uniformly valid inference. They assume that the estimator used to construct the counterfactual is asymptotically normally distributed. However, as explained in Appendix B.4, with few treated clusters, asymptotics have not kicked in, and no Central Limit Theorem can be applied. This is why, in the following analysis, the estimator proposed by Manski and Pepper (2013, 2018) is preferred. Another reason to choose bounds over the approaches proposed by Rambachan and Roth (2023) is that the former offers a valid alternative to the TWM, as it is robust to violations of both PT and NA assumptions.[26]

7 Robustness Checks

Regressions (2) and (3) are only two possible version of the TWM approach. Much more heterogeneity than that introduced in (2) and (3) could be per se allowed by exploiting the estimator proposed by Wooldridge. For instance, the ATT may be allowed to vary not only over time but also across treatment cohorts (e.g., see regression (B.10)). Besides allowing the ATT to vary over time and across treatment cohorts, the treatment dummy can be interacted with controls, cohorts, and year dummies altogether (for example, see Eq. (B.11) in Appendix B). Thus allowing much more flexibility and, consequently, relaxing the PT and NA assumptions much more than in (3).

However, since in such a context, the number of observations is fixed and small, whereas T is large, rather than allowing the treatment effects to vary across treated cohorts and over time and then aggregating them ex-post in some way – as suggested by Callaway and Sant'Anna (2021) – ex-ante restrictions on the τ's have been imposed in (2) and (3). Specifically, the ATT's were allowed only to vary over time and covariates, where covariates were centered about their mean conditionally on being treated only. This latter choice was because of efficiency reasons, as explained in Subsection 6.1.

[26] As a robustness check, the approach proposed by Rambachan and Roth (2023) is also used. However, if $M = 0$ is imposed – the absence of violations of the PT in the post-treatment period – the confidence intervals computed using this approach are much larger than those obtained for the same parameter using TWM or the estimator proposed by Sun and Abraham (2021). This corroborates the idea that normality is not likely to hold in this scenario.

In the following analysis, another version of the TWM will also be estimated as a robustness check. Specifically, the $ATTs$ will be allowed to vary by calendar time, but homogeneity across cohorts will be imposed. If, on the one hand, PdRs are region-specific, there is no reason to expect that regions treated in different cohorts experience different TEs.[27] Conversely, the impact of RPs likely varies over time.[28] The calendar-type versions of the TWM that will be estimated are the following:

$$y_{g,t} = \alpha + \sum_{s=2007}^{2018} \tau_s(w_{g,t} \cdot fs_t) + \sum_r \lambda_r d_{g,r} + \eta_t + u_{g,t} \tag{4}$$

$$\begin{aligned} y_{g,t} =& \alpha + \sum_{s=2007}^{2018} \tau_s(w_{g,t} \cdot fs_t) + \sum_r \lambda_r d_{g,r} + \sum_r (d_{g,r} \cdot \mathbf{x}_g)\boldsymbol{\gamma}_r + \mathbf{x}_g\boldsymbol{\kappa} \\ &+ (w_{g,t} \cdot \dot{\mathbf{x}}_{g,treat})\boldsymbol{\rho}_{treat} + \eta_t + u_{g,t} \end{aligned} \tag{5}$$

where the only difference with (2) and (3) lies in the fact that $w_{g,t}$ is now interacted with fs_t (and not with $intens_{g,t}(s)$), and fs_t represents a dummy variable equal to 1 if $s = t$ and zero otherwise. Compared to (2) and (3), $fs_t = 1$ tells how much being treated in a specific year (e.g., 2008) affects the dependent variable of interest. Whereas, in (2), the estimated τ tells how much being s periods away from when the region first signed a PdR impacts y. Please note that this version of the TWM is closer to bounds, as the ATT will be allowed to vary by calendar time and not by how many periods had passed from when the region first received the treatment.

As far as standard errors are concerned, as anticipated in Subsection 6.1, all regressions' coefficients will be reported with bootstrap confidence intervals – where the bootstrap procedure used is the sub-cluster wild bootstrap proposed by MacKinnon and Webb (2018). However, standard errors will also be estimated using the classic CRVE as a robustness check.

The TWM is just one of the different approaches that have been recently proposed to estimate the ATT consistently in a context with variation in treatment timing (Appendix B.3). To check whether the results obtained via the TWM approach are robust, the estimator proposed by Sun and Abraham (2021) will also be used. Specifically, this method allows us to obtain an estimator for τ_s similar to the one obtained by estimating Eq. (2) and (3) via POLS. To do so, Sun and Abraham (2021) exploit a generalization of the PT and NA assumptions similar to the one required for the validity of the TWM. However, compared to the estimator proposed by Wooldridge (2021), it has one main disadvantage. The TE is not explicitly allowed to vary with covariates, making the parallel trend and no-anticipation assumptions less likely to be valid.[29]

For what concerns bounds, instead, as anticipated in Subsection 6.3, the largest values of δ observed in the pre-treatment periods will be used. This is because the larger the uncertainty allowed, the more credible the results (Manski, 2003). However, to understand whether the results are robust, bounds will 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 built invoking the SC or time-invariance assumptions. Not only, as an ulterior sensitivity exercise, a complete set of combinations of

[27] As an ulterior sensitivity exercise, the ATT was also allowed to vary by cohorts, but no clear pattern was detected across cohorts. However, this result could be entirely driven by the fact that few treated regions exist in the second and third cohorts, causing standard errors to be imprecisely estimated.

[28] As pointed out in Section 3, for example, the first version of RPs mainly focused on cost-containment. Whereas, from the second round of PdRs, the law gave ELAs and financial stability equal importance.

[29] For further details on how this estimator works, please refer to Sun and Abraham (2021).

δ_{SC} and δ_{time} will also be employed. 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 TE is not identified – this implies that we can reject the hypothesis the time invariance and SC assumptions jointly hold in the data.

As pointed out in Subsection 6.2, while it is reasonable that regions that first received the treatment in 2007 could not forecast the introduction of *Piani di Rientro*, it is hard to believe that later treated regions did not anticipate their entrance in a PdR. Despite bounds allowing to retrieve a consistent estimator of the TE without knowing the exact reason for which the invariance assumption (needed for identification) does not hold, different sensitivity exercises will be run to understand whether it is plausible that the (C)NA assumption holds.

First, following Wooldridge (2021), leads of the treatment dummy will be introduced in Eq. (4) and (5). Failing to reject the null hypothesis of joint insignificance of the coefficients for the treatment dummy leads will indicate that the (C)NA assumption is likely to hold. However, as for the PT assumption, conditioning the analysis on passing a test would generate a pre-test bias (Roth et al., 2023). For this reason, in the spirit of the estimator proposed by Callaway and Sant'Anna (2021), Eq. (4) and (5) will be re-estimated by anticipating for each region the entrance in PdR by two years (e.g., Abruzzo will first join a PdR in 2005). The choice to anticipate the treatment by two years is because a first draft of Recovery Plans was outlined with the budget law for 2005, but this measure has never become effective.

Lastly, following Leive et al. (2023) and McKibbin (2023), who exclude from their analysis later-treated units for which they expect the NA to be potentially violated, all the analyses described in Subsection 6.1 will be re-run by excluding later-treated regions (that is, Calabria, Piemonte, and Puglia). Specifically, the following two regressions will be estimated via POLS:

$$y_{g,t} = \alpha + \sum_{s=0}^{11} \tau_s(w_{g,t} \cdot intens_{g,t}(s)) + \lambda d_{g,2007} + \eta_t + u_{g,t} \tag{6}$$

$$\begin{aligned} y_{g,t} =& \alpha + \lambda d_{g,2007} + \sum_{s=0}^{11} \tau_s(w_{g,t} \cdot intens_{g,t}(s)) + \mathbf{x}_g\boldsymbol{\kappa} + (d_{g,2007} \cdot \mathbf{x}_g)\boldsymbol{\gamma} + \\ & (w_{g,t} \cdot \dot{\mathbf{x}}_{g,treat})\boldsymbol{\rho}_{treat} + \eta_t + u_{g,t} \end{aligned} \tag{7}$$

Furthermore, calendar-type versions of the TWM, like the ones reported in (4) and (5), will also be estimated. To save space, these latter results are available upon request.

8 Results

As explained in Section 3, *Piani di Rientro* aimed at restoring the financial stability of RHSs while guaranteeing (or even improving) ELAs. These two targets, however, are incoherent unless efficiency gains are feasible. While there is wide consensus in the existing literature that RPs managed to reduce budget imbalances (e.g., Arcà et al. (2020); Bordignon et al. (2020)), on the other side, there is contrasting evidence regarding the impact on efficiency and the quality of services provided (Aimone Gigio et al., 2018; Depalo, 2019; Bordignon et al., 2020; Guccio et al., 2023; Beraldo et al., 2023; Cirulli and Marini,

2023). In particular, there is also a disagreement between the Central Government and treated regions regarding the impact of such a policy on health outcomes (SiVeAS, 2014; Calabrò, 2016; Aimone Gigio et al., 2018).

This section will provide novel evidence regarding the causal impact of *Piani di Rientro* on costs, inefficiency, and quality indicators. Specifically, compared to existing studies, this paper will consider explicitly the staggered nature of the policy and the possibility of having heterogeneous treatment effects across regions and over time. To corroborate the choice to rely on methods that explicitly model that the policy exhibits variation in treatment timing, I will also exploit the DiD decomposition proposed by Goodman-Bacon (2021).

Before showing the results obtained through the DiD Decomposition, I will present some descriptive statistics on the covariates used in the following analysis.

Next, the results obtained through the estimator proposed by Wooldridge (2021) will be presented. As pointed out in Subsection 6.1, this method will allow us to retrieve a consistent estimator of the ATT in a context with variation in treatment timing by introducing treatment effect heterogeneity. Moreover, this estimator will also estimate the policy's long-run impact.

Lastly, I will show how the results obtained through the TWM approach are robust to the relaxation of the main identifying assumptions. To do so, building on Depalo (2019), the non-parametric estimator proposed by Manski and Pepper (2018) will be used. Specifically, compared to Depalo (2019), the impact of Recovery Plans will be evaluated not only on regions that first signed a PdR in 2007 but also on later treated regions (Calabria, Piemonte, and Puglia). In addition, results for each post-treatment year will be provided rather than showing the effect only at the end of the first and second rounds. Thus, we can estimate the long-term consequences of *Piani di Rientro*.

8.1 Descriptive Statistics – Covariates

Table 2 reports descriptive statistics for all the covariates described above split by treatment cohort and by before and after 2007, the first year in which an RP was signed. The choice to split the sample only into a pre and after-2007 analysis was for the sake of brevity.[30] On average, regions that adopted PdRs in different cohorts share similar characteristics (in terms of population distribution), which have remained stable over time. Conversely, huge differences arise between treated cohorts and never treated regions, especially regarding GDP per capita.[31]

8.2 Results – DiD Decomposition

Most existing studies on RPs exploit some variation of the classic TWFE estimator to evaluate the impact of the policy on different health indicators. However, recent studies in the DiD literature show that failing to take into account the staggered nature of the policy might invalidate inference (see Appendix B.2 for a review). To show why it is important to consider this aspect of *Piani di Rientro*, the DiD decomposition will be employed in this subsection. Only the results for the mortality rate from ischaemic heart diseases will be shown to save space. The intuition behind the results obtained for this indicator is similar to that for the other variables considered (available upon request).

[30] Further descriptive statistics are available upon request.

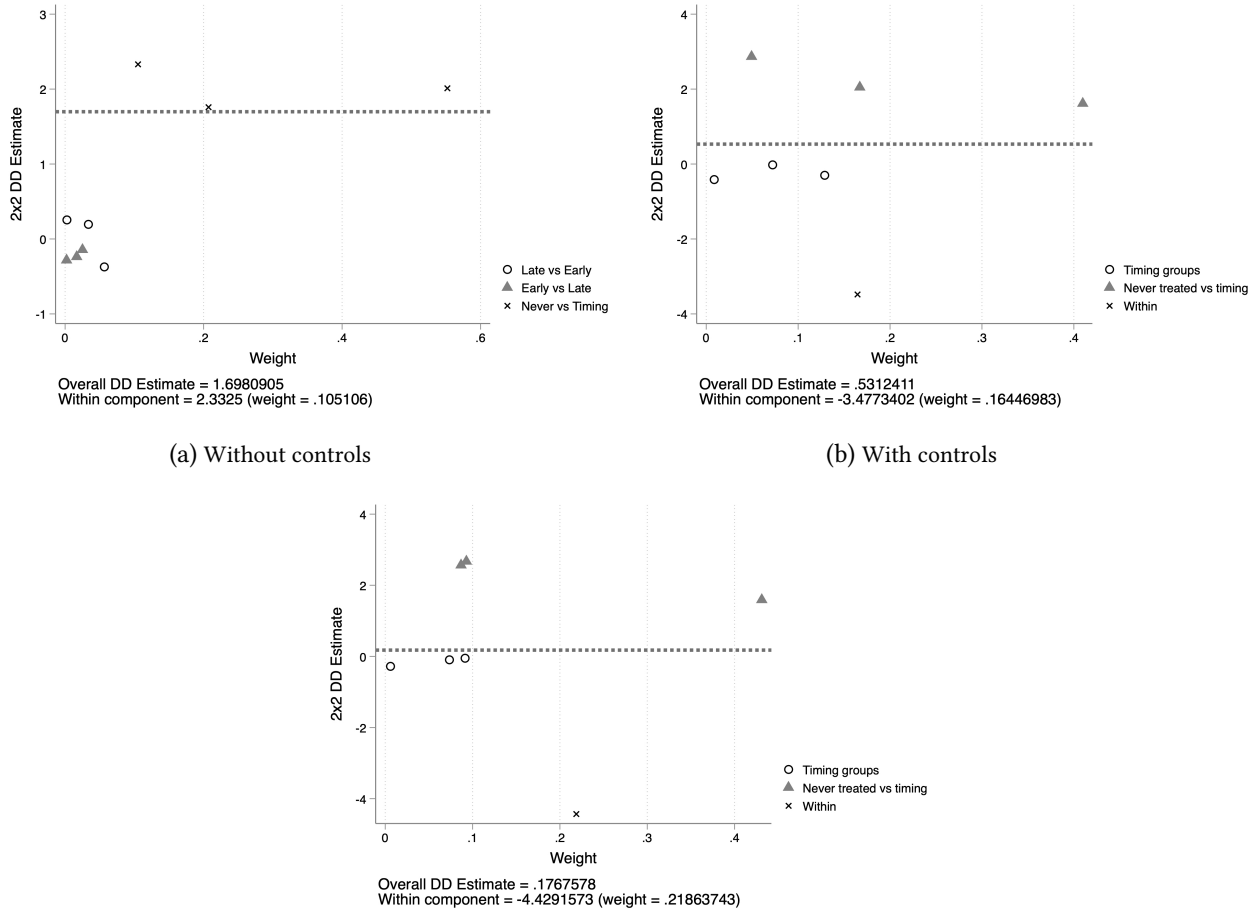
[31] Please note that the TWM allows centering $\mathbf{x}$ about the mean of the $\mathbf{x}$ over the treated cohorts. However, since no huge differences were detected in observable characteristics among treated cohorts, the x's were centered about their mean conditionally on being treated for efficiency reasons.

Table 2: Descriptive statistics

	Pre 2007				Post 2007			
	Never treated	**d2007**	**d2010**	**d2011**	**Never treated**	**d2007**	**d2010**	**d2011**
Population size	2401136.359	2956629.413	3128209.667	4037581.556	2572743.803	3048436.131	3172304.167	4094854.083
	(2614284.400)	(2091093.773)	(1151858.005)	(11962.393)	(2828880.410)	(2193615.634)	(1241479.770)	(84077.806)
% people aged 15-34	25.314	26.730	26.498	29.141	20.956	22.402	22.353	24.145
	(2.010)	(2.925)	(2.862)	(1.490)	(1.710)	(2.626)	(2.939)	(1.464)
% people aged 35-54	28.889	28.129	28.110	27.273	30.530	29.987	29.516	29.319
	(1.148)	(1.238)	(1.273)	(0.713)	(0.855)	(1.085)	(0.894)	(0.424)
% people aged 55-64	12.317	11.568	11.800	10.906	12.653	12.839	12.866	12.468
	(0.913)	(1.250)	(1.568)	(0.476)	(0.709)	(0.813)	(0.681)	(0.452)
% people aged ≥ 65	20.033	19.035	19.289	16.121	22.139	21.452	21.827	19.683
	(2.315)	(3.606)	(2.324)	(0.935)	(2.038)	(3.251)	(2.393)	(1.468)
% female	51.383	51.643	51.386	51.444	51.376	51.570	51.415	51.479
	(0.392)	(0.567)	(0.315)	(0.037)	(0.378)	(0.502)	(0.239)	(0.047)
GDP per capita	26615.164	19986.604	19720.645	15052.733	31334.810	23425.171	22926.078	17315.515
	(5194.182)	(5455.010)	(6166.943)	(1358.625)	(6072.568)	(5801.506)	(6365.046)	(739.143)

Notes: Means are reported with s.d. in parentheses. **Never-treated** = Regions that have never undergone an RP. **d2007** = Regions that first signed an RP in 2007 (Lazio, Abruzzo, Campania, Liguria, Molise, Sardegna, and Sicilia). **d2010** = Regions that first signed an RP in 2010 (Calabria and Piemonte). **d2011** = Regions that first signed an RP in 2011 (Puglia).

Figure 3: DiD Decomposition – Mortality rate from ischaemic heart diseases



(a) Without controls

(b) With controls

(c) Restricted sample + controls

In Panel (a) of Figure 3 is plotted each pair of (2×2) DiD estimators against their weights, obtained by applying the DiD decomposition to the estimated coefficient for β in (1), using as dependent variable the mortality rate from ischaemic heart diseases.[32] In particular, the crosses represent terms in which never-treated regions act as controls and regions treated at some point act as treated. The open circles are terms in which 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 in which early treated (e.g., those treated in 2007) units act as controls for later treated regions (those joining an RP in 2010 or 2011), known in the literature 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 gauge the extent of the bias by visually inspecting which of the three types of groups (*Early vs. Late*, *Late vs. Early*, *Never vs. Timing*) receive more weights. Summing each comparison group's weights tells how much of the identifying variation comes from a specific group, thus helping to understand how well the TWFE estimator performs compared 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 means that the estimated coefficient for β in (1) is (slightly) biased downward.

Panel (b), instead, reports the DiD decomposition applied to a variation of (1) in which population size, its related distribution by age and gender, and GDP per capita are also controlled for. 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 new source of variation comes from the fact that time-varying controls have now been included in (1). In particular, this latter term takes into account the fact that units belonging to the same timing group may feature different values of the covariates. Also, in this case, almost 63% of the overall identifying variation comes from *Never vs. Timing* terms, with an average effect of $1.83p.p.$. 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 timing terms, for an average effect of $-.20$. Also, in this case, the bias drives down the estimated coefficient.

Lastly, since the DiD decomposition assumes that treated units are at an absorbing state, Panel (c) reports the same results shown in Panel (b), but where units that left the treatment status – Liguria, Piemonte, and Sardegna – have been dropped. Despite the overall coefficient being much smaller than that reported in Panel (b), it is still evident that the estimated coefficient is biased downward by timing groups.

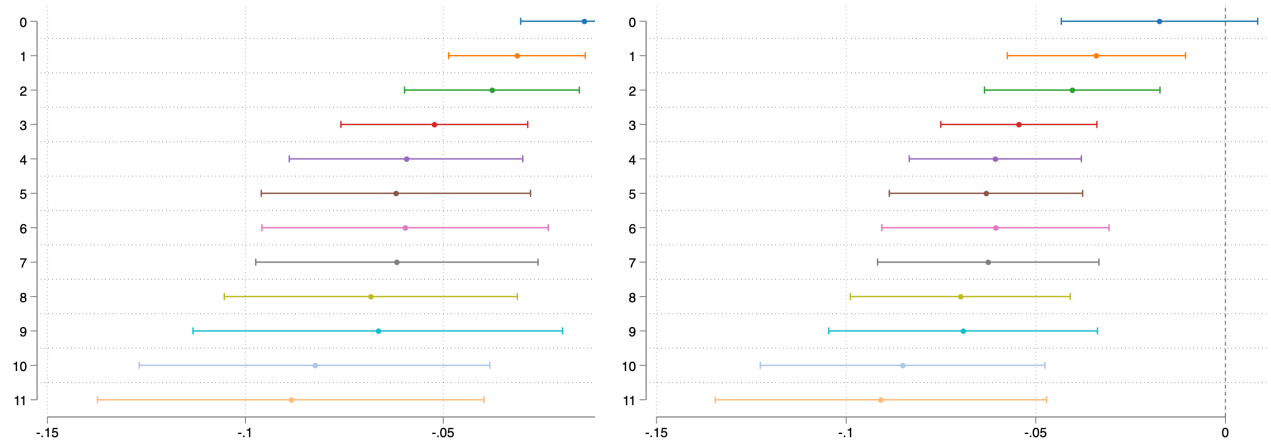
These graphs show that the estimated coefficient obtained through the classic TWFE estimator is biased in such a context. Therefore, as already pointed out, ignoring the staggered nature of the policy would jeopardize inference. To prevent this problem, in the next subsection, results obtained by an approach that allows the researcher to retrieve a consistent estimator of the ATT in a context with variation in treatment timing are shown.

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

8.3 Results – TWM

This subsection will present results obtained using the Two-Way Mundlak approach. As mentioned in Section 6, this method allows us to consistently estimate the ATT in a context with variation in treatment timing.

Figure 4: TWM results – 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 (2) using as the dependent variable the log of current health expenditure. In Panel (b) are reported the results obtained by estimating (3). 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.

In Panels (a) and (b) of Figure 4, the estimated coefficients for $intens(s)$ obtained by estimating Eq. (2) and (3), respectively, using as the dependent variable the log of current health expenditure, are depicted. On the x-axis, points estimates are reported, while on the y-axis, how many periods at time t had passed from when region g first received the treatment.[33] Coefficients are reported with 95% confidence intervals, obtained via the subcluster wild bootstrap with Rademacher weights (MacKinnon and Webb, 2018). This bootstrap procedure requires that the t-statistic be obtained through a CRVE estimator (Appendix B.4). In contrast, the resampling is carried out at a finer level than the level at which the disturbances are clustered. Specifically, in this context, standard errors and, consequently, the t-statistic are clustered at the RHS level, and the resampling is carried out at the RHS-year level.[34]

On average, with the introduction of RPs, the Central Government managed to reduce current health expenditure in treated regions (as expected). Except for the coefficient for $intens(0)$, the effect is always negative and statistically significant at 5%, and this result holds for both two Panels. One reason for failing to reject the null hypothesis that the coefficient for $intens(0)$ is equal to 0 could be because it may take time for regions to adjust to the PdR-specific requirements and, thus, be able to reduce current health expenditure.

[33] For instance, $intens(0)$ means that 0 periods had passed from when the region first received the treatment (for Abruzzo, which was first treated in 2007, $intens(0)$ will be equal to 1 in 2007 while, for Calabria, it will be equal to 1 in 2010). Whereas $intens(5)$ means that 5 years had passed from when the region first received the treatment (thus, for Abruzzo, it will take value 1 in 2012).

[34] Similar results were obtained using standard errors obtained via a classic CRVE estimator. These results are shown in Figure C.2 in Appendix C .

If, on one side, the fact that results are stable across the two specifications is reassuring. On the other side, what also emerges by comparing Panel (a) with Panel (b) is that conditioning or not on covariates only marginally affects the results (the estimated coefficients are similar in magnitude). This latter finding could be because the PT and NA assumptions are not completely relaxed by controlling only for very few covariates.

Before showing the results for the other indicators, it is worth dedicating a few words about the bootstrap confidence intervals. Three assumptions need to be satisfied for the subcluster wild bootstrap to perform well (Appendix B.4). The first assumption – equal cluster sizes and 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 to be satisfied in this context, as any cross-cluster heteroskedasticity is allowed by this assumption. The second assumption – the average within-cluster correlation needs to be small – on the other side, is more difficult to check in practice. MacKinnon and Webb (2018) show that, however, one way to check whether this assumption holds is to inspect whether the p-values of the restricted and unrestricted bootstrap are similar, and in particular, whether the p-value for the restricted subcluster wild bootstrap is always larger than the one of the unrestricted. To check whether this condition holds in this context, in Figure 5, the bootstrap distribution of the t-statistic for each value of $intens(s)$ is reported by imposing and not imposing the null hypothesis. To save space, only the results for the coefficients for $intens(s)$ obtained by estimating (3) – using as dependent variable the log of current health expenditure – are shown.[35] What emerges is that the bootstrap p-values of the restricted and unrestricted subcluster wild bootstrap are very close, suggesting that assumption 2 should not be seriously violated.

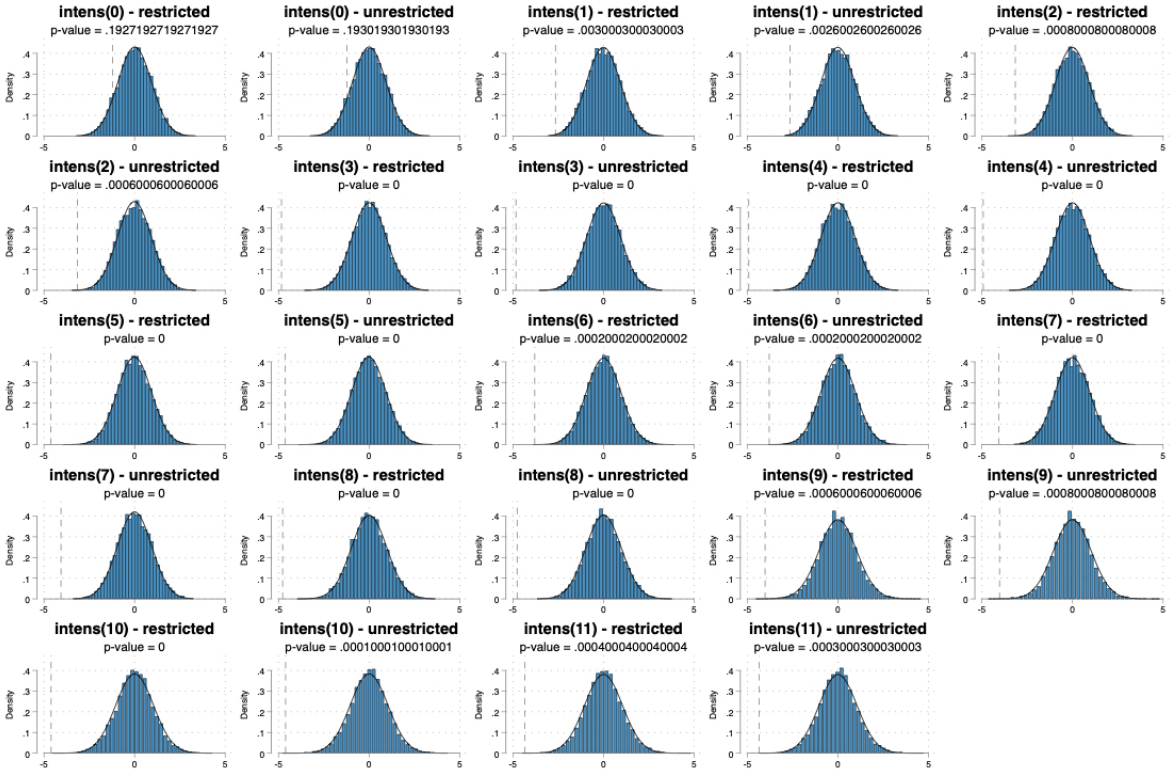
Results for the other indicators discussed in Section 5 are reported in Figure 6. To save space, results obtained only by conditioning on covariates are reported (i.e., by estimating (3)), as the intuition is similar to that for the log of current health expenditure – conditioning on covariates only marginally affects the $ATTs$. As in Figure 4, also in this latter case, coefficients are reported with 95% confidence intervals obtained via the subcluster wild bootstrap, which proved to work well for all these indicators (assumption 2 should not be seriously violated in any of these cases). The bootstrap distributions of the t-statistic for each value of $intens(s)$ for each indicator are available upon request.

For what concerns the two targets of the PdRs, we can see that, as expected, the introduction of Recovery Plans reduced the hospitalization rate both for total and also for acute hospitalizations (Panel (a) and (b), respectively) and the number of hospital beds (Panel (c)) in treated regions. For each dependent variable, coefficients are always statistically different from 0 except for the first two years (and for $intens(11)$ in the case of hospital beds). The absence of an effect in the years immediately after the implementation of the plans is coherent with the fact that the policy requires a structural reorganization of the RHS. Therefore, it may take time for the region to meet the requirements adequately.

As far as inefficiency is concerned (Panel (d) in Figure 6), it seems that after the introduction of PdRs, treated regions experienced a small but statistically significant reduction in the level of inefficiencies as proxied by the ratio between the hospitalization for DRG at high risk of inappropriateness versus those that could not be avoided. The lower the ratio, the lower the level of inefficiency in the RHS. Also, in this case, coefficients for $intens(s)$ are almost everywhere statistically significant at the 5%. On the other side, the coefficients for $intens(s)$ are always indistinguishable from 0 if the c-section

[35] Similar results were obtained for the coefficients of (2), which are available upon request.

Figure 5: Bootstrap distribution t-statistics – ln(Current health exp)



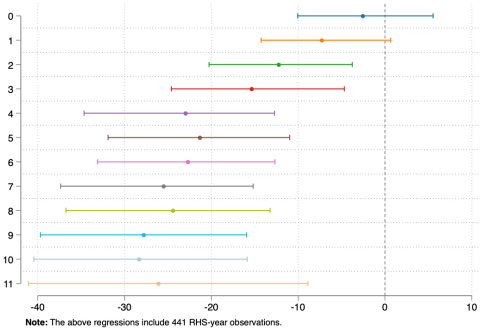
Notes: The above histograms report the bootstrap distribution of the t-statistic for each of the coefficients for $intens(s)$ obtained by estimating (3) using as dependent variable the log of current health expenditure. In particular, these bootstrap distributions were obtained by exploiting the subcluster 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.

rate (Panel (e)) is considered. However, the lack of an effect for this latter indicator could be because the estimator for the ATT is inconsistent. Indeed, what emerges from Figure 2 is that the parallel trends assumption is unlikely to be valid for this indicator. On the contrary, it is likely to be satisfied concerning the inefficiency ratio. If this is true, then PdRs may have effectively reduced inefficiencies.

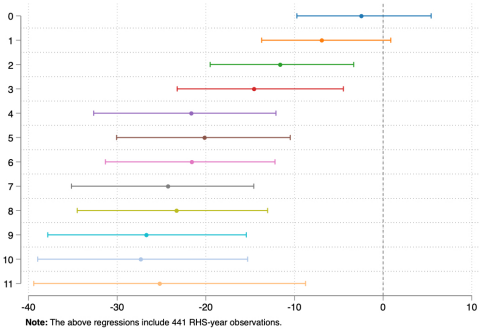
If, on the one side, PdRs managed to reduce costs, on the other side, they seem to have caused a deterioration in the quality of services provided. Indeed, except for the $\%$ patients with hip fracture replacement (Panel (h)), a deterioration in the other quality indicators is detected. Before commenting on the results for the other quality indicators, it is worth understanding why no effect is detected for the $\%$ patients with hip fracture replacement. There could be two potential explanations. First, from the contextual perspective, this indicator is part of the *specific targets* of PdRs. Thus, regions and hospitals have an economic incentive to ensure that patients hospitalized with a hip fracture undergo surgery within 48 hours from when they arrive in the emergency room. Second, from the econometric side, as it is evident from Figure 2, the parallel trends assumption is likely to be violated for this indicator. In addition, the CPT is also likely not to hold since only a few control variables were included. Therefore, it could be that the absence of an effect is due to the estimator of the ATT being inconsistent.

For what concerns the other proxies of quality, what emerges is that the introduction of PdRs has led to an increase in the mortality rate (all causes) – as can be seen in Panel (f) – and the mortality rate from

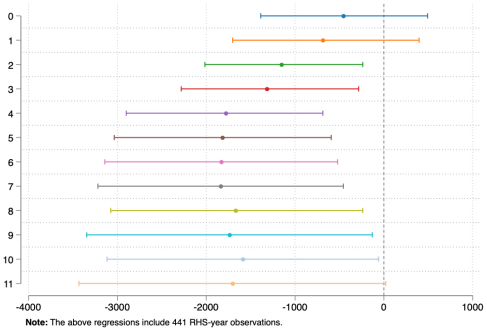
Figure 6: TWM results – Other indicators



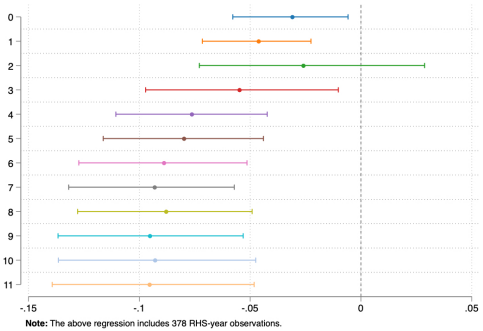
(a) Hospitalization rate



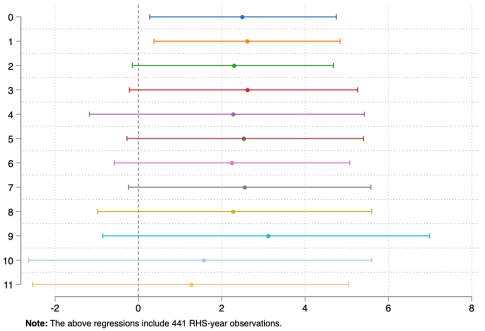
(b) Hospitalization rate – acute



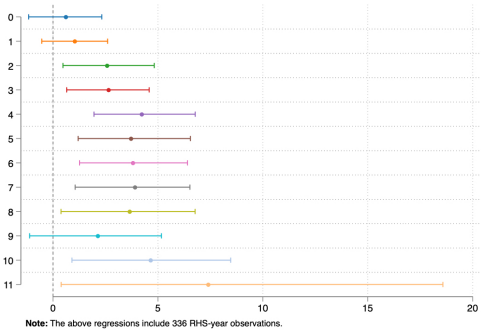
(c) Hospital beds (ordinary regime)



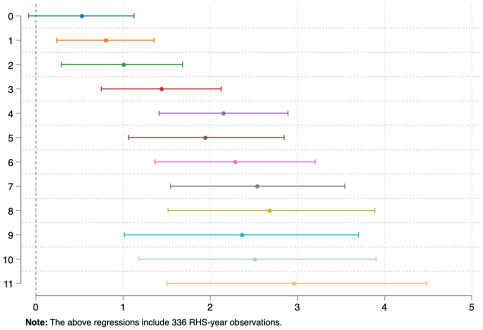
(d) Inefficiency ratio



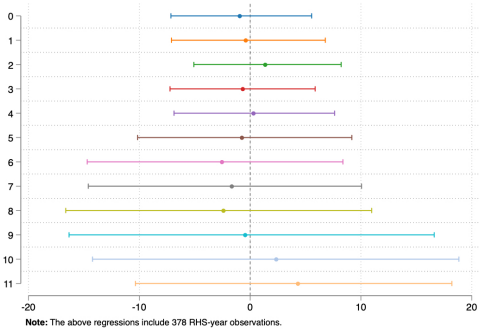
(e) % C-sections



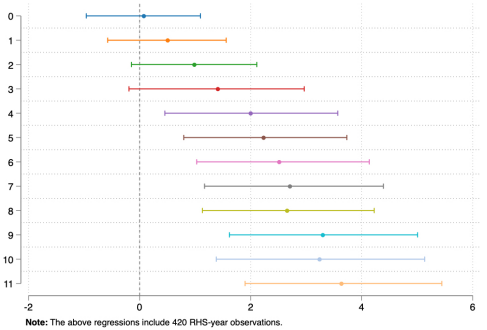
(f) Mortality rate



(g) Mortality rate from ischaemic heart diseases



(h) % patients with hip fracture replacement



(i) % patients migrating for the ordinary acute hospit.

Notes: Results depicted in Panel (a) to (i) were obtained by estimating (3) 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.

ischaemic heart diseases (Panel (g)). In particular, for the latter, the coefficients for $intens(s)$ obtained by estimating (3) are statistically different from 0 almost everywhere, with point estimates ranging from barely 1 to almost 3 p.p. The increase in mortality rate (all causes) documented in treated regions after the introduction of PdRs is in line with existing studies documenting that healthcare spending cuts are associated with increased mortality (e.g., Golinelli et al. (2017)). On the other hand, the increase in AMI mortality rate is likely to have been caused either by the reduced budget available to hospitals (Kwok et al., 2018; Lobo et al., 2020) or by a decrease in hospital supply Brown et al. (2014).

Regarding the impact of the policy on the % of patients migrating for ordinary acute hospitalizations (Panel (i)), a statistically significant increase is documented for regions being 4 periods away or more from when they had first signed an RP. The increase documented for this indicator in treated regions aligns perfectly with the results obtained by Beraldo et al. (2023) as well as with existing studies reporting that patients' mobility flow rises following healthcare spending cuts (Arcà et al., 2020) and reduced hospital supply (Balia et al., 2018). The coefficients for $intens(0)$ to $intens(3)$ being statistically indistinguishable from 0 could be because, as explained above, it may take time for the region to adjust to the PdR's specific targets. In addition, people may take additional time to perceive a variation in service quality. While the rise documented in the mortality rates (all causes and from ischaemic heart diseases) may be a direct negative consequence of the policy, the increase in the mobility flow could be explained by the perceived decreased quality of services provided in the patient's region of origin.

8.4 Results – Bounds

As explained in Section 6, the TWM approach requires three strong assumptions to be valid for the estimator for the ATT to be consistent. Besides the fact that the treatment needs to be irreversible (i.e., no leavers), the TWM hinges upon a generalization of the PT and NA assumptions to the multiperiod setup with variation in treatment timing for identification. In this context, however, these assumptions are likely violated. First, as explained in Section 3, three regions left the treatment status over the observed period. Second, as illustrated in Figure 2, the PT assumption is unlikely to hold for (at least) some indicators. Third, as pointed out in Subsection 8.3, point estimates are only marginally affected when conditioning on covariates. This may indicate that introducing these covariates is insufficient to "credibly" relax the parallel trends. Lastly, while it is reasonable to assume that regions that first signed an RP in 2007 could not forecast the introduction of such a policy, it is hard to rule out anticipation behaviors for regions that entered the treatment status in subsequent years. If any of the three identifying assumptions is seriously violated, then the estimator for the ATT can be proved to be inconsistent.

One way to overcome the abovementioned issues is to give up "exact" invariance assumptions in favor of milder ones. Specifically, building on Depalo (2019), results obtained by the estimator proposed by Manski and Pepper (2018) will be presented hereafter. There are three main advantages of using bounds. First, the uncertainty about the validity of the exact invariance assumption can be directly taken into account, and, at the same time, the reason for this latter not holding should not be known to the researcher. The higher the level of uncertainty permitted, the more credible the results will be. Second, exploiting this non-parametric estimator allows estimating region-specific treatment effects, $\tau_{g,r,t}$. This can help to shed light on the results found using the TWM approach. Third, bounds allow combining more invariance assumptions. This translates into a refinement of the interval length in which $\tau_{g,r,t}$ should lie and prevents the researcher from favoring any particular assumption over the

other.

Following Depalo (2019), time invariance and synthetic control assumptions will be exploited jointly in the subsequent analysis. The covariates used to construct the benchmark employing the SCM are always those described in Section 5. However, given the excess variability that characterizes each of the indicators employed in the analysis, a time-demeaning is applied before employing any invariance assumptions to compute bounds on $\tau_{g,r,t}$. In particular, the mean computed before the first PdR was signed (i.e., 2007) is subtracted from each dependent variable. The choice to rely on the mean computed before 2007 rules out any anticipatory effects of later-treated regions. Therefore, bounds will be computed using – rather than $y_{g,t} - y_{g,t} - \bar{y}_g$ where $\bar{y}_g = \frac{1}{T}\sum_{t=y^{first}}^{2006} y_{g,t}$ and $T = (2006 - y^{first} + 1)$.

Table C1 of Appendix C reports, for each (treated) region, the difference between the observed outcome and estimated counterfactual built invoking either SC or time invariance assumptions is reported until the last year before the introduction of the RP. 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 benchmark obtained by relying on one of the two invariance assumptions (Manski and Pepper, 2018; Depalo, 2019). The reason to favor a larger value of δ is that the more uncertainty is permitted, the more credible the results. If, for a given indicator, the largest (absolute) difference amounts to .1 or less, the specific δ is increased by .1 for all regions.

In Table 3 the bounds on the $\tau_{g,r,t}$ based on the optimal values of δ (i.e., $\delta_{T,Max}$, $\delta_{SC,Max}$) for costs indicators are shown. Those for the TE for efficiency indicators are reported, instead, in Table 4, and for quality's proxies in Table 5. Specifically, results for each treated region are reported in the columns, while results for each post-treatment period are shown in the rows. Before proceeding with commenting on the findings for these indicators, it is worth highlighting that missing values observed for Calabria and Piemonte for the years 2007-2010 and for Puglia for the years 2007-2011 – for each indicator – are because these regions entered the treatment status in 2010 and 2011, respectively. If any other missing value is observed, this is because the upper bound is lower than the lower bound. For instance, considering the results obtained for hospital beds, we can see four missing values for Liguria. One in 2014, and the other three missing values for the years 2016-2018. These missing values are caused by the fact that in these four years for Liguria, the estimated upper bound for $\tau_{g,r,t}$ is lower than the lower bound.

For what concerns results obtained using the log of current health expenditure as the dependent variable, surprisingly, there seems to be no effect. Except for Liguria (years 2016–2018) and Sicilia (in 2011), for which both the upper and the lower bounds are below zero – meaning that the TE is negative –, bounds for all the other regions cover 0. In these cases, it is impossible to say anything about the TE. On the other side, there seems to be a positive effect for Sardegna (years: 2012, 2014–2018). This is counterintuitive, as a decrease in health expenditure should be observed after introducing a PdR. However, Sardegna is assumed to be at an absorbing state, but it left the treatment status in 2010, invoking its special statute. Similar reasoning applies for Liguria and Piemonte – that left the treatment status, having reached the plan's goals. As stated before, one of the advantages of the bounds is that it allows estimating a treatment effect separately for each treated region. Therefore one can ignore the columns for Liguria, Piemonte, and Sardegna if one thinks that assuming that these regions are at an absorbing state is too demanding.

Overall, the absence of an effect on the log of current health expenditure could be explained by two reasons. First, as it is evident in Figure 2, with the introduction of RPs, healthcare spending did not

Table 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 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

decrease for treated cohorts but only flattened. Furthermore, the trend experienced by treated regions parallels that of never-treated RHS. Second, in Figure C.1 are reported the estimated coefficients of τ_s obtained by estimating Eq. (5) using as dependent variable the log of current health expenditure (this version of the TWM is the one closer to bounds). Even though a negative and statistically significant effect is detected, point estimates are very small in magnitude. Therefore, it could be that the absence of an effect is because the ATT, which was already small, was split further into ten different region-specific treatment effects.

As far as the impact of Recovery Plans on hospital beds is concerned, what emerges from Table 3 is that a negative effect is documented for almost all of the treated regions. This result is expected, as this was one of the targets of the Recovery Plans. In particular, the bounds for the TE are both negative for Abruzzo (only the last two years), Calabria, Lazio (only from 2009), Molise (from 2008), Sardegna (from 2010), Sicilia (from 2007), Piemonte and Puglia (only in 2017 and 2018). By contrast, the effect of Liguria is sometimes estimated to be positive. Again, this puzzling result is probably because Liguria exited the RP.

For what concerns hospitalization rate, as expected, the bounds for the treatment effects are almost in each year and for each region negative – implying that the treatment effect is negative. The bounds cover 0 in almost all the post-treatment years for Liguria and Piemonte, and this finding could be because these regions left the treatment status over time. By contrast, the upper bound for Lazio is almost everywhere lower than the lower bound. Similar results were found for the hospitalization rate built by considering acute hospitalizations only, which are shown in Table C2.

In Table 4, bounds for the ratio between the hospitalization for DRG at high risk of inappropriateness versus those that could not be avoided are shown. Except for Abruzzo (from 2010) and Calabria (from 2014), there appear to be no gains in terms of efficiency in treated regions after the introduction of the policy. By contrast, in Campania and Sardegna (2009–2011) and Molise (2009–2010), the inefficiency levels seem to have deteriorated.

The fact that the estimated impact reported in Table 4 is different from the one obtained using the TWM approach (depicted in Figure 6) could be because at least one of the main identifying assumptions required for the TWM to yield a consistent estimator of the ATT being seriously violated. On the other side, it could also be the case that the results shown in Figure 6 – which were very small – were entirely driven by efficiency gains in Calabria and Abruzzo. If, instead, one considers the estimated coefficients obtained by estimating (5), the null hypothesis of the insignificance of the coefficient is never rejected (results available upon request).

On the other hand, PdRs seem to have had a positive effect on the c-section rate. What emerges from Table 4 is that, except in some cases where the estimated upper bound is lower than the lower bound, the TE is almost everywhere positive. If, on one side, this latter result may indicate that after the introduction of RPs, the appropriateness of care provided deteriorated. On the other hand, as pointed out in Section 5, since c-sections are larger in private hospitals (De Luca et al., 2021), this finding may be indicative of a declined perceived quality of healthcare services provided by public hospitals, rather than public hospitals effectively providing less appropriate care. Again, the fact that the estimated TE is different from the one obtained using the TWM approach could be because at least one of the main identifying assumptions required for the TWM to yield a consistent estimator of the ATT is seriously violated.

For what concerns quality indicators, in Table 5 results for the mortality rate from ischaemic heart

Table 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						

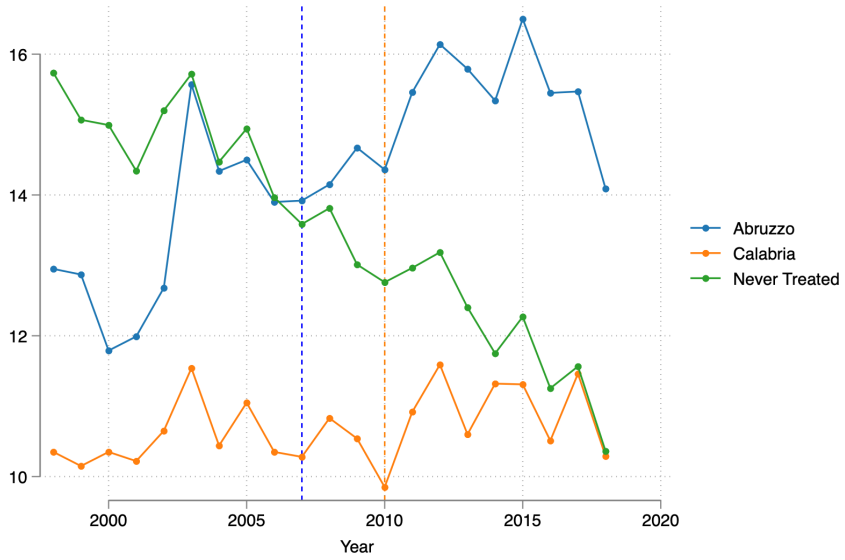
Table 5: Bounds on treatment effect based on ($\delta_{T,Max}$, $\delta_{SC,Max}$)

	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

diseases and the % of patients migrating to other regions for ordinary acute hospitalizations are reported. As far as the former indicator is concerned, an increase in the mortality rate from ischaemic heart diseases is observed in Abruzzo (from 2009 to 2017), Campania (only for four years), Liguria (from 2011 to 2016), Molise (almost in each year). For these regions, the lower and the upper bound are almost in every case above 0. By contrast, no clear patterns are detected for other treated regions. Similar results were obtained using the mortality rate (all causes). However, when considering this latter indicator, the fraction of cells in which the upper bound is lower than the lower bound is larger than the case in which the mortality rate from ischaemic heart diseases is considered. Results for this latter indicator are reported in Table C2.

The absence of an effect of RPs on the mortality rate from ischaemic heart diseases for Calabria could be due to an identification problem rather than the absence of an effect. Indeed, Calabria has always had the lowest mortality rate from ischaemic heart diseases. To try to understand this puzzling result, Figure 7 depicts the time-series for the mortality rate from ischaemic heart diseases for Calabria against that for Abruzzo and the average for the never-treated regions. What is clear from this picture is that the small increase detected in Calabria after the introduction of the RP might have been offset by the large decrease experienced by never-treated regions. On the other side, this indicator was increasing for Abruzzo. After 2007 (when the region first signed an RP), it continued rising, with the gap between its time-series evolution and that of never-tread growing over time.

Figure 7: 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.

For what concerns the impact of the policy on the % of patients seeking care outside the region of origin, what emerges from Table 4, is 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 taken cautiously, as by looking only at the % of patients, the spatial dimension of the process is completely ignored. Moreover, the % is computed over all regions. However, it is well known in Italy that the flow is not bidirectional (Balia et al., 2018; Arcà et al., 2020). Most of the flow comes from the Southern –

and, at the same time, poorest – regions of Italy seeking care in the Center and North of Italy hospitals.

As pointed out by Arcà et al. (2020), the fact that an impact is detected could indicate that the introduction of the policy has hampered the gap in the equality of access between the poor and the rich, besides amplifying the existing gap between the South and the Center-North of Italy in terms of quality of services provided.

Lastly, in Table C2, results for the % of patients with hip fracture replacements are shown. What emerges is that the upper bound is lower than the lower bound in almost every post-treatment year and for each region. Thus, the TE is not identified. There could be two potential explanations for failing to find an effect for this latter indicator. First, this indicator is part of the ex-post monitoring. Therefore both hospitals and regions have economic incentives to monitor this indicator and ensure that adequate levels of this target variable are reached. Second, the lack of an effect could also be due to an identification problem since, in most cases, the upper bound is lower than the lower bound.

8.5 Robustness checks

In this subsection, the results obtained by carrying out the different sensitivity exercises discussed in Section 7 will be presented.

As discussed in Section 7, Eq. (2) and (3) are only two possible version of the TWM approach. In theory, more heterogeneity could be allowed (e.g., by letting the ATT vary by treatment cohort and over time). However, for efficiency reasons, ex-ante restrictions have been imposed to estimate Eq. (2) and (3).

Another version of the TWM is the one presented in Eq. (4) and (5), where the ATT have been allowed to vary by calendar time. Figure C.1 of Appendix C plots $\hat{\tau}_s$ obtained by estimating (5) via POLS 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. While along the x-axis, point estimates are shown as before. On the y-axis, the estimated coefficients for $w_{g,t}fs_t$ are reported. For instance, 2007 will be a dummy equal to 1 if the region is treated in 2007 (this will be equal to 1 for Abruzzo, but not for Piemonte, which was treated for the first time in 2010).

Results shown in Figure C.1, are in line with those presented in Figures 4 and 6. The estimated effect on the mortality rate is always positive and statistically significant, whereas the impact on current health expenditure is still negative. The only difference with results reported in Figures 4 and 6 lies in the first three estimated coefficients (those for 2007, 2008, and 2009), which are now not statistically different from 0. However, the absence of a statistically significant effect for these coefficients could be because, in these three years, only seven regions were treated. In contrast, in 2010, Calabria and Piemonte joined a PdR, and from 2011, also Puglia. Indeed, in the case in which the ATT is let vary by calendar time, as in Eq. (5), the interpretation of the coefficients is no longer in terms of how many periods had passed from the first year in which the region was first treated. Rather, the $\hat{\tau}_s$'s tell the effect for regions effectively treated in year $t = s$. To give the intuition, $\hat{\tau}_{2015}$ will tell how much being treated in 2015 affects the dependent variable under scrutiny. Results obtained by estimating (5) using as dependent variable the other indicators are available upon request.

Next, as an ulterior sensitivity exercise, standard errors of Eq. (2) to (5) were re-estimated using a classic cluster-robust variance-covariance estimator. Figure C.2 mirrors Figure 4, where now standard

errors have been obtained via a standard CRVE. What emerges is that results depicted in Figure C.2 are perfectly in line with those plotted in Figure 4. The estimated τ_s's are always statistically different from 0 at 5%, except for the coefficient of $intens(0)$, as in Figure 4. However, confidence bands are now slightly wider than before. This finding perfectly aligns with Monte Carlo simulations results in MacKinnon and Webb (2018), suggesting that the classic CRVE will likely perform poorly in such a context.

Then, to understand whether the results obtained via the TWM approach are robust, the estimator proposed by Sun and Abraham (2021) was also used. Similar results to those depicted in Figures 4 and 6 were obtained by exploiting this latter estimator. To save space, only results for the log of current health expenditure are shown in Figure C.3, whereas those obtained using the other indicators as regressand are available upon request. What emerges from Figure C.3 is that the point estimates are similar to those obtained by estimating (3) via POLS. However, three things are worth stressing. First, the estimator proposed by Sun and Abraham (2021) does not allow for the relaxation of the PT and NA assumptions, whereas in (3) both these two assumptions were relaxed by controlling for covariates. Second, inspecting the estimated coefficients for the intensity indicators in the pre-treatment periods reveals that some are statistically different from zero. This will test whether the PT assumption will likely hold in practice. In particular, for the PT to be likely to hold, one should fail to reject the null hypothesis of joint insignificance of all the coefficients for the intensity indicators in the pre-treatment period (i.e., $s < 0$). If one rejects the null hypothesis of joint insignificance, then pre-trends may be present. However, as pointed out by Roth (2022), these tests may suffer from several issues and should be used cautiously.[36] Third, the estimated coefficients reported in Figure C.3 are reported with 95% confidence intervals, where standard errors were estimated using a standard CRVE (the level at which s.e. were clustered is at the RHS level). The STATA package `eventstudyinteract` provided by Sun and Abraham (2021), indeed, does not allow to resort to any bootstrap procedure to overcome the issue of having few treated clusters (see, Appendix B.4 for further details). Overall, inference obtained via this estimator may be highly unreliable.

As far as bounds are concerned, δ_{SC} and δ_{time} were set equal to the largest absolute difference observed in the pre-treatment period between the actual outcome and the one built invoking the respective invariance assumption to increase the credibility of the results. However, to understand whether results are robust, bounds 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}). Results based on $\delta_{SC,p_{75}}$ and $\delta_{time,p_{75}}$ are reported in Table C3. To save space, only results for the TE for the mortality rate from AMI are shown (results for the other indicators are available upon request). What emerges from Table C3 is that, despite the fraction of bounds not identified is slightly larger than in Table 5, results are similar. Overall, it seems that, after the introduction of RPs, there was an increase in the mortality rate from ischaemic heart diseases.

Not only were bounds re-estimated by setting the δ's equal to p_{75}, but as an ulterior sensitivity exercise, a complete set of combinations of δ_{SC} and δ_{time} was also be employed. To save space, only results for the mortality rate from ischaemic heart diseases for Abruzzo are shown in Table C4. Overall, results are in line with those reported in Table 5 for Abruzzo, and this holds for different combinations of δ_{SC} and δ_{time} (especially from 2014). Interestingly, when setting jointly δ_{SC} and δ_{time} equal to 0,

[36] For further details on why these tests may be misleading, please refer to Section 6.

the upper bound is always smaller than the lower bound, suggesting that the time invariance and SC assumptions never hold jointly in the data. This latter result is perfectly in line with what was found by Depalo (2019). Similar findings to those shown in Table 5 using δ_{MAX} were obtained for the other regions exploiting different combinations of δ_{SC} and δ_{time}, indicating that results obtained via bounds are robust.

Next, different sensitivity exercises were performed to check whether the (C)NA assumption would likely be satisfied. First, Eq. (4) and (5) were re-estimated by including up to three leads of the treatment dummy to test for the presence of anticipatory effects. Regardless of the dependent variable considered, I always 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, Eq. (2) to (5) were also re-estimated by anticipating for each region the entrance in PdR by two years. For the sake of brevity, these results are not shown. What emerges is that the coefficients for $intens(0)$, $intens(1)$, f_{2005}, and f_{2006} are always not statistically different from 0, suggesting that there should not be anticipatory effects. These results hold for each dependent variable, except the hospitalization rate, for which the coefficients are positive and statistically different from 0.

Lastly, results obtained by re-estimating Eq. (6) and (7) by excluding later-treated regions aligns perfectly with those shown in Subsection 8.3. These results are available upon request.

Overall, these findings indicate that, except for the c-section rate, anticipation behaviors should be a minor issue in such a context.

8.6 Discussion

Overall, results obtained using the TWM approach align with those found using milder non-parametric assumptions. The ATT's estimated relying on the method proposed by Wooldridge (2021) go in the same direction as the TE's obtained via Manski and Pepper (2018)'s non-parametric estimator, and this is true for each of the indicators under scrutiny, except for the log of current health expenditure, the c-section rate, and the inefficiency ratio. In addition, results are robust to different sensitivity exercises (see Subsection 8.5), suggesting that there may not be serious violations of the main identifying assumptions needed for the parametric estimator to be consistent.

In light of the findings in the previous subsections, there is evidence that with the introduction of *Piani di Rientro*, the Central Government effectively reduced costs and current health expenditure in treated regions. In particular, cost containment was achieved through reduced hospital beds and hospitalization rates (as expected by the policymaker). As also pointed out by Depalo (2019), reducing hospitalization rates is the easiest way to reduce healthcare spending since they account for approximately 90% of the current health expenditure of Italian RHSs.

For what concerns current health expenditure – as explained in the previous subsections – the lack of an effect when considering bounds could be due to identification problems rather than the absence of such an effect. This hypothesis can be corroborated by the fact that after 2007, a substantial reduction in the total deficit for public health spending is documented, as shown in Figure 1.

Although from the second version of *Piani di Rientro*, the Central Government gave both cost-reduction and ELAs the same importance, these two objectives are incoherent unless the RHS can experience efficiency gains. Using as a proxy of inefficient use of resources the ratio between the hospitalization for DRG at high-risk inappropriateness versus those that cannot be avoided – one of

the indicators constituting the *Griglia LEA* and part of the ex-post monitoring – what emerges is that, except for Abruzzo and Calabria, the introduction of RPs did not reduce the level of inefficiencies in treated RHSs. Moreover, when considering as a proxy for the inappropriateness of care the % c-sections, a positive effect is detected; 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 effectively performing more c-sections.

On the other hand, a deterioration in the quality of services provided – as proxied by the mortality rate from ischaemic heart diseases – and in the perceived quality – as proxied by the % of patients migrating to other regions for acute hospitalizations – is observed in treated regions (also in the long run). Furthermore, introducing RPs seems to have had a negative – but to some extent smaller – effect on the mortality rate (all causes) in treated RHSs. The fact that all of these indicators appear to have been affected by the policy is well in line with the existing studies documenting that the mortality rate, and in particular those from ischaemic heart diseases, is sensitive to variations in hospital supply and healthcare spending cuts may have adverse (negative) effects (Brown et al., 2014; Golinelli et al., 2017; Kwok et al., 2018; Arcà et al., 2020; Lobo et al., 2020). Also, the fact that the policy influences patients' mobility is not surprising, as one of the main predictors of interregional patients' mobility in Italy is hospital supply (Balia et al., 2018).

Lastly, rank correlation coefficients were employed to understand whether the increase in the mortality rate from ischaemic heart diseases and patients' mobility was larger in regions that have experienced a larger reduction in hospital beds and hospitalization rates. Specifically, Kendall's τ and Spearman's ρ were considered.

Both the two 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 TE for hospitalization rate and that of patients' mobility is also negative. The larger the reduction in the hospitalization rate, the worse the consequences on quality indicators are in line with existing studies in medicine and health economics considering hospitalization rate as a proxy for quality rather than for 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.

Overall, these findings suggest that, with the introduction of RPs, the Central Government effectively reduced healthcare spending and costs. Nonetheless, contrary to what was expected from the policymaker, the policy did not improve the efficiency of RHSs, either in the long run, except for Abruzzo and Calabria. By contrast, regions that experienced the largest reduction in hospitalization rates were also those for which the largest negative consequences on the quality of the RHS were detected. These findings are in line with those found by Depalo (2019). However, compared to this latter study, having a much longer post-treatment period allows us to test whether the quality of services provided by the RHSs is restored in the long run.

9 Conclusions

This paper contributes to the ongoing debate in health economics regarding the impact of healthcare spending cuts on citizens' health outcomes. To this aim, novel evidence on the effect of a cost-containment measure first introduced in Italy in 2007 on the performance of Italian RHSs is provided.

Analyzing the Italian NHS is interesting since it is a decentralized-based, almost entirely publicly funded system, where regions retain RHS's management power. Despite the decentralization of the NHS guarantees, to a certain extent, autonomy to local governments in the organization of the RHS, from the 2000s, Essential Levels of Assistance were introduced to ensure that each citizen was granted a minimum level of services and that these were equivalent across regions, thus allowing comparability across Italian RHSs in terms of service quality and efficiency. Nonetheless, regions exhibit substantial variability in the quality of services provided (Aimone Gigio et al., 2018), with the Northern regions providing, on average, higher quality services.

Although, over the years, different measures have been implemented to devolve more responsibilities to regional governments in managing the RHS to contain costs, the Central Government has continued to finance ex-post the large deficit run by regional governments. This has led the public expenditure, and consequently, the total deficit for public health spending, to grow dramatically. In this context, in 2007, the Central Government introduced *Piani di Rientro* to curb excessive spending and restore the financial stability of regional governments.

Sixteen years after the first PdR was signed, there is a unanimous consensus that this policy managed to cut costs. However, most existing studies find contrasting evidence regarding the impact of the policy on citizens' health outcomes and the efficiency of RHSs.

This study provides novel evidence of the policy's causal impact on a set of quality, efficiency, and cost indicators. In particular, besides considering indicators that are part of the ex-post monitoring, other commonly-used quality proxies are used. By analyzing the impact of the policy on variables that are not part of the ex-post monitoring, whether regions have strategically outperformed on indicators that discriminate on their ability to receive funds can be tested.

To the best of my knowledge, this is the first paper that assesses PdR's long-run impact, considers the staggered nature of the policy, and allows heterogeneity of treatment effects across regions and over time. By resorting to the DiD decomposition (Goodman-Bacon, 2021), why failing to take into account the staggered nature of the policy and the fact that the treatment effect is likely to vary over time and across treated RHSs may lead to biased results is shown. Then, to consider the features of the policy and overcome the pitfalls of the classic TWFE estimator – commonly used in the literature of PdRs – the estimator proposed by Wooldridge (2021) is relied upon. This approach allows retrieving a consistent estimator of the ATT in a context with variation in treatment timing by introducing heterogeneity of treatment effects within the simple linear regression framework. Moreover, compared to existing studies, to overcome the problem of having few (treated) clusters – which leads the classic cluster-robust variance-covariance estimator to over-reject H_0 – the subcluster wild bootstrap proposed by MacKinnon and Webb (2018) is also employed.

However, the Two-way Mundlak approach, besides requiring that the treatment is at an absorbing state (i.e., no leavers), hinges upon a generalization of the parallel trends and no-anticipation assumptions for identification. If any of these assumptions are seriously violated, the treatment effect's estimator can be proved inconsistent. Since, in this context, at least one of these assumptions is likely violated for some dependent variables, results obtained by relying on milder non-parametric assumptions are also provided, the bounded variation assumptions proposed by Manski and Pepper (2018).

Using bounds allows the researcher to directly model the uncertainty about the validity of identifying assumptions to retrieve a consistent estimator of the TE. In addition, one of the main advantages of relying on assumptions of bounded variation is that the exact reasons for violating the identifying

assumption should not be known to the researcher, with the only requirement being that the level of uncertainty is specified ex-ante. The higher the level of uncertainty, the more credible the results Manski (2003). Besides technical reasons, exploiting this non-parametric approach is interesting as it makes it possible to estimate region-specific treatment effects, thus shedding light on the results obtained using the parametric approach.

Overall, the results obtained using the above estimators suggest that PdRs effectively reduced costs, and cost containment was pursued by reduced hospital beds and hospitalization rates. However, contrary to what was expected from the policymaker, except for Abruzzo and Calabria, the policy did not improve the efficiency of treated RHSs. Still, it led to an unintended deterioration in the quality of treated regions – as proxied by indicators not part of the ex-post monitoring. These findings are robust to different sensitivity exercises and hold even in the long run. For almost all the variables, results obtained via the parametric approach go in the same direction as those estimated using bounds. I view the fact of performing inference relying on both the parametric (TWM) and the non-parametric approach as one of the strengths of this paper, as this allows us to understand how the results are robust to the relaxation of the identifying assumptions, as also stated by Manski and Pepper (2018).

Lastly, the larger negative consequences were documented in those regions that experienced a more drastic reduction in hospitalization rates. This last finding aligns with existing studies documenting that decreasing the hospitalization rate is associated with worse health outcomes (e.g., Berchialla et al. (2010)).

To conclude, the results presented in this paper suggest that Recovery Plans effectively reduced costs. Nonetheless, cost containment did not translate into efficiency gains but an unintended deterioration in the quality of services provided by treated RHSs. I believe these findings, coupled with the fact that regions that faced a less abrupt reduction of costs experienced a less severe deterioration in the quality of services provided, may inform policymakers about the importance of taking a more gradual approach toward healthcare spending cuts. Sudden drops in health spending unavoidably translate into worse health outcomes without necessarily enhancing the efficiency of healthcare providers.

References

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

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

L. Aimone Gigio, D. Alampi, S. Camussi, G. Ciaccio, P. Guaitini, M. Lozzi, A. L. Mancini, E. Panicara, and M. Paolicelli. 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, 03 2018.

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

M. Andrews, O. Elamin, A. R. Hall, K. Kyriakoulis, and M. Sutton. 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, 2017.

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

E. Arcà, F. Principe, and E. Van Doorslaer. Death by austerity? The impact of cost containment on avoidable mortality in Italy. *Health Economics*, 29(12):1500–1516, 2020. doi: https://doi.org/10.1002/hec.4147. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hec.4147.

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

S. Beraldo, M. Collaro, and I. Marino. Patient migration as a response to the regulation of subnational healthcare budgets. *Regional Studies*, 0(0):1–13, 2023. URL https://doi.org/10.1080/00343404.2023.2171379.

P. Berchialla, F. Foltran, A. Di Lenarda, G. Barbati, L. Franchin, L. Salvatore, D. Gregori, et al. 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, 2010.

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

M. Bordignon, S. Coretti, M. Piacenza, and G. Turati. Hardening subnational budget constraints via administrative subordination: The Italian experience of recovery plans in regional health services. *Health Economics*, 29(11):1378–1399, 2020. doi: https://doi.org/10.1002/hec.4138. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hec.4138.

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

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

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

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

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

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

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

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

A. Cook, M. Gaynor, M. Stephens Jr, and L. Taylor. The effect of a hospital nurse staffing mandate on patient health outcomes: Evidence from California's minimum staffing regulation. *Journal of health economics*, 31(2):340–348, 2012.

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

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

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

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

Eurostat. European system of accounts - ESA 2010, 2013. URL https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-02-13-269.

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

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

D. Golinelli, F. Toscano, A. Bucci, J. Lenzi, M. P. Fantini, N. Nante, and G. Messina. 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, 2017.

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

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

C. Guccio, G. Pignataro, D. Romeo, and F. Vidoli. 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, Nov. 2022. URL https://ideas.repec.org/p/yor/hectdg/22-28.html.

C. Guccio, G. Pignatora, and F. Vidoli. 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, Mar. 2023. URL https://ideas.repec.org/p/yor/hectdg/23-02.html.

R. Heijink, X. Koolman, and G. Westert. 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, June 2013. doi: 10.1007/s10198-012-0398-3. URL https://ideas.repec.org/a/spr/eujhec/v14y2013i3p527-538.html.

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

K. Imai and I. S. Kim. On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data. *Political Analysis*, 29(3):405–415, 2021. doi: 10.1017/pan.2020.33.

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

Institute of Medicine. *Crossing the Quality Chasm: A New Health System for the 21st Century*. The National Academies Press, Washington, DC, 2001. ISBN 978-0-309-46561-8. doi: 10.17226/10027. URL https://nap.nationalacademies.org/catalog/10027/crossing-the-quality-chasm-a-new-health-system-for-the.

C. S. Kwok, S. V. Rao, J. E. Potts, E. Kontopantelis, M. Rashid, T. Kinnaird, N. Curzen, J. Nolan, R. Bagur, and M. A. Mamas. 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, 2018.

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

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

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

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

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

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

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

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

C. F. Manski and J. V. Pepper. 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, May 2018. URL https://ideas.repec.org/a/tpr/restat/v100y2018i2p232-244.html.

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

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

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

MoH. Analisi spazio-temporale degli indicatori della Griglia LEA relativi all'adempimento sul "mantenimento dell'erogazione dei LEA" - Trend regionali anni 2001 - 2012, 2014. URL https://www.salute.gov.it/portale/lea/dettaglioPubblicazioniLea.jsp?lingua=italiano&id=2245.

MoH. Monitoraggio dei LEA attraverso la cd. Griglia LEA, 2015. URL https://www.salute.gov.it/portale/lea/archivioPubblicazioniLea.jsp?lingua=italiano&area=Lea&pubblicazioni.page=1.

MoH. Monitoraggio dei LEA attraverso la cd. Griglia LEA, 2016. URL https://www.salute.gov.it/portale/lea/archivioPubblicazioniLea.jsp?lingua=italiano&area=Lea&pubblicazioni.page=1.

MoH. Monitoraggio dei LEA attraverso la cd. Griglia LEA, 2017. URL https://www.salute.gov.it/portale/lea/archivioPubblicazioniLea.jsp?lingua=italiano&area=Lea&pubblicazioni.page=1.

MoH. Monitoraggio dei LEA attraverso la cd. Griglia LEA, 2018. URL https://www.salute.gov.it/portale/lea/archivioPubblicazioniLea.jsp?lingua=italiano&area=Lea&pubblicazioni.page=1.

MoH. Rapporto annuale sull'attività di ricovero ospedaliero - Dati SDO 2020, 2020.

L. Moja, A. Piatti, V. Pecoraro, C. Ricci, G. Virgili, G. Salanti, L. Germagnoli, A. Liberati, and G. Banfi. 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. 2012.

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

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

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

OECD. *Health at a Glance 2019*. 2019. doi: https://doi.org/https://doi.org/10.1787/4dd50c09-en. URL https://www.oecd-ilibrary.org/content/publication/4dd50c09-en.

OECD. *Health at a Glance 2021*. 2021. doi: https://doi.org/https://doi.org/10.1787/ae3016b9-en. URL https://www.oecd-ilibrary.org/content/publication/ae3016b9-en.

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

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

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

C. Pross, R. Busse, and A. Geissler. 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, 2017.

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

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

J. Roth. Pretest with caution: Event-study estimates after testing for parallel trends. *American Economic Review: Insights*, 4(3):305–22, September 2022. doi: 10.1257/aeri.20210236. URL https://www.aeaweb.org/articles?id=10.1257/aeri.20210236.

J. Roth, P. H. Sant'Anna, A. Bilinski, and J. Poe. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, 2023. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2023.03.008. URL https://www.sciencedirect.com/science/article/pii/S0304407623001318.

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

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

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

F. Schiele, C. P. Gale, E. Bonnefoy, F. Capuano, M. J. Claeys, N. Danchin, K. A. Fox, K. Huber, Z. Iakobishvili, M. Lettino, et al. 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, 2017.

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

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

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

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

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

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

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

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

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

A Appendix A

Table A1: 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 $17+$)

B Causal Inference Review

In this Appendix, a short overview of how to derive the counterfactual outcome – that is, the outcome one would observe had the policy not been implemented –and, in turn, the average treatment effect on the treated will be presented. In particular, in Subsection B.1, which are the main assumptions made in the program evaluation literature, and what these assumptions imply will be reviewed. Next, Subsection B.2 will show why the classic Two-way Fixed Effects estimator – which is commonly used in the existing literature to estimate the impact of Recovery Plans on health-related outcomes – might be biased in a context with variation in treatment timing, as the one under analysis. Subsection B.3 will then present the estimator proposed by Wooldridge (2021), which overcomes the issues discussed in Subsection B.2 and allows the researcher to retrieve a consistent estimator of the causal estimand of interest in a staggered treatment adoption setup. Lastly, Subsection B.4 will briefly review the problem of how to perform (valid) inference in a context with few treated clusters.

B.1 Deriving the Counterfactual Outcome

Following Neyman (1923), Rubin (1974), and others, the potential outcome framework is used to define the parameters of interest. For simplicity, the same notation introduced in Section 6 will be employed. 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, $t = 1, \ldots, T$.[37]

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 $te_{g,t} = y_{g,t}(r) - y_{g,t}(\infty)$.[38] Then, the realized outcome for a generic unit can be rewritten as:

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

where $d_{g,q} + d_{g,q+1} + \ldots + d_{g,T} + d_{g,\text{never}} = 1$. Oftentimes, however, the causal estimand of interest

[37] For simplicity, to keep notation in line with that introduced in Section 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).

[38] 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 $r + 1$.

is the average treatment effect on the treated, ATT, in periods where the treated cohorts are effectively under the policy. That is:

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

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,[39] 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 $pre \in \{1, \dots, q\}$ 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.[40] 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 $\mathbb{E}(y_t(\infty) - y_1(\infty)|d_q, \dots, d_T) = \mathbb{E}(y_t(\infty) - y_1(\infty))$).[41] 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).

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

[40] 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 B.3.

[41] 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.

B.2 TWFE & DiD Decomposition

As mentioned in Section 6, most of the existing studies on recovery plans estimate the causal effect of the policy relying on some variation of the classic TWFE estimator. However, there exists a recent literature in program evaluation proving that the estimated coefficient of the treatment dummy obtained via TWFE is an inconsistent estimator for the causal estimand of interest in a context with variation in treatment timing (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:

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

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 $k \in t = 2, \dots, T$ (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.[42] As shown in Figure B.1, 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:

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

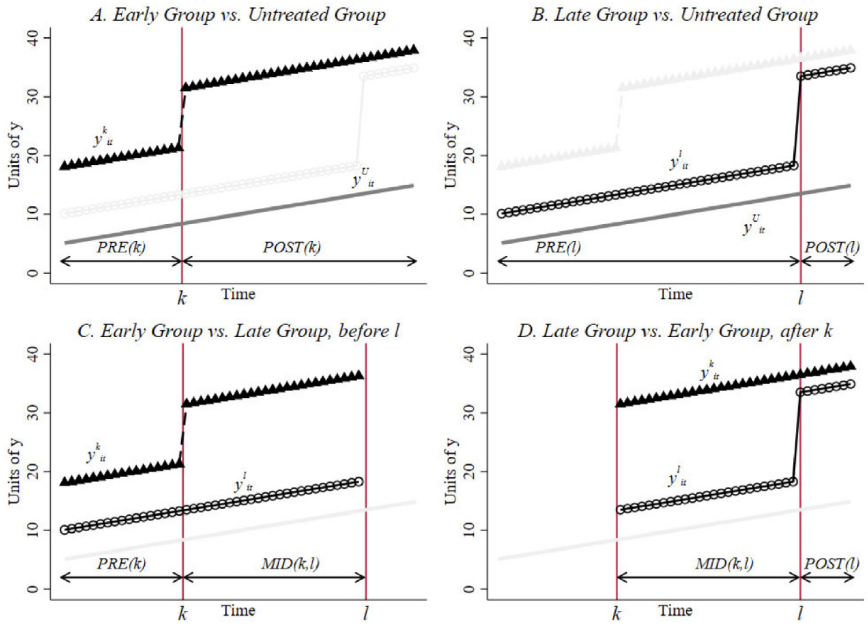
Panels C and D of Figure B.1, instead, plot what would happen if Eq. (B.2) is estimated using only timing groups (that is, groups k and l). Identification in these two plots comes from these two groups receiving the treatment in different periods. Specifically, before l, the early-treated group act as 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

[42] 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 B.1: 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).

B.3 Two-Way Mundlak Approach (TWM)

Different estimators have been recently proposed to prevent the issue of negative weights discussed in Section B.2 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).[43]

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).[44] 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 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

[43] 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.

[44] 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), 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).[45] 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.

B.4 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.[46] However, when G is small, the cluster-robust t-statistic can severely over-reject

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

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

(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:

$$y_{g,t} = \mathbf{z}_{g,t}\boldsymbol{\theta} + u_{g,t} \tag{B.12}$$

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 (B.12) 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, 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.

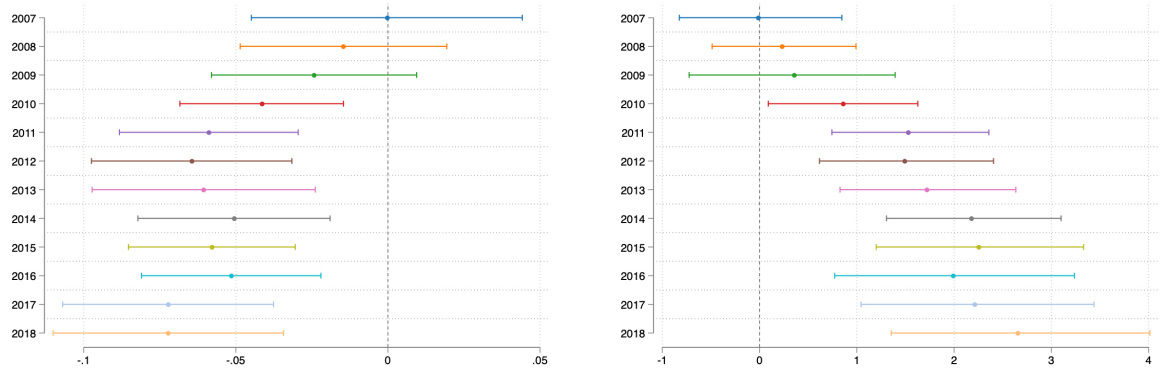
2. The average within-cluster correlation should be small when the number of treated clusters is small.

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).

C Additional Results

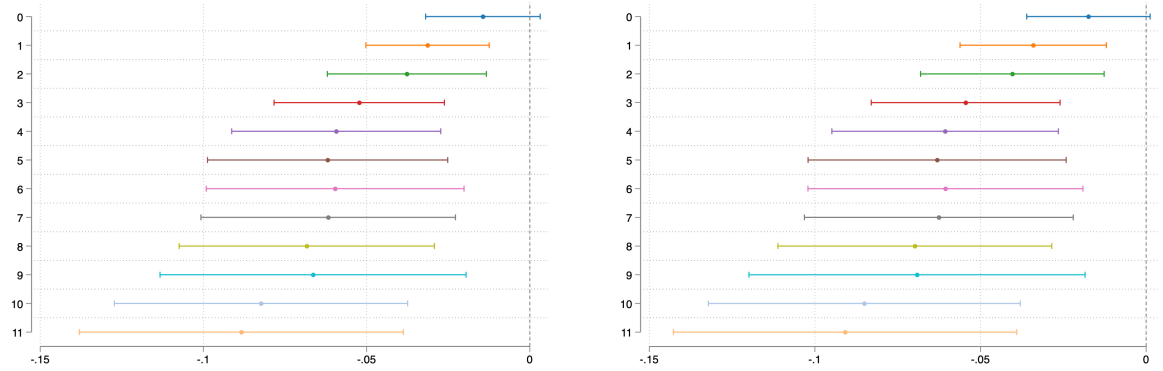
Figure C.1: TWM results – ATT that varies by calendar time



(a) ln(Current health exp) (b) Mortality rate from ischaemic heart diseases

Notes: Results depicted in Panel (a) were obtained by estimating (5) using as the dependent variable the log of current health expenditure. Panel (b) reports $\hat{\tau}_s$ obtained by estimating (5) 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.

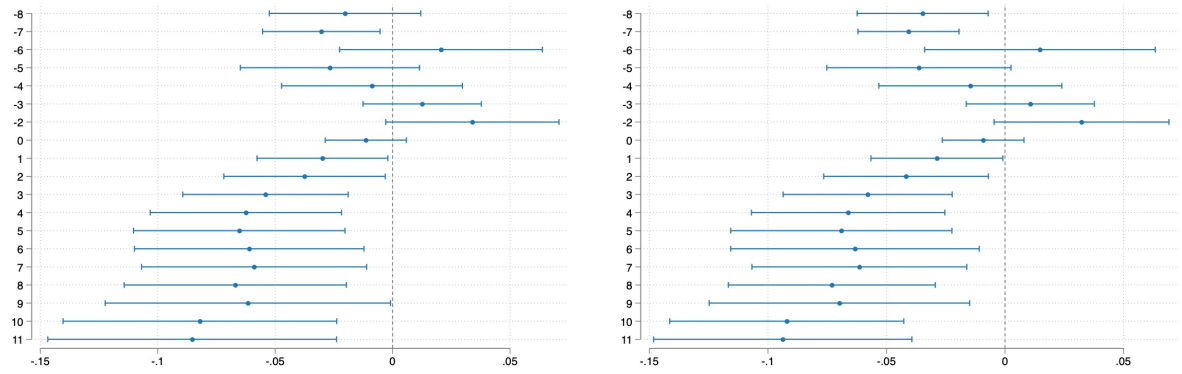
Figure C.2: TWM results – ln(Current health exp) with cluster-robust s.e.



(a) Without controls (b) With controls

Notes: The above regressions include 357 RHS-year observations. Results depicted in Panel (a) were obtained by estimating (2) using as the dependent variable the log of current health expenditure. In Panel (b) are shown the results obtained by estimating (3). 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 C.3: Sun and Abraham results – $ln(\text{Current health exp})$



(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.

Table C1: 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 C1 (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 C1 (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

(Continues)

Table C1 (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}
2001		-0.0		0.0		-0.0		0.0		0.0		0.0		0.0		-0.0		-0.0		0.0
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

(Continues)

Table C1 (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}
2007			1.8	4.9									0.5	0.9	0.7	5.6				
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

(Continues)

Table C1 (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}
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
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				

(Continues)

Table C1 (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	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
	% 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 C2: 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 ≤ 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 C3: 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 C4: 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

(Continues)

Table C4 (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	4.0	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
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

(Continues)

Table C4 (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	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
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

(Continues)

Table C4 (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	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
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

(Continues)

Table C4 (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	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
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