CIRANO
Allier savoir et décision

NETWORKS AND QUEUES: EVIDENCE FROM MEDICAL REFERRALS IN CANADA



VINCENT BOUCHER
JOSETTE GBETO
MARIPIER ISABELLE

CS
2026s-27
CAHIER SCIENTIFIQUE

CIRANO
Allier savoir et décision

Les **cahiers de la série scientifique** visent à rendre accessibles les résultats des recherches effectuées par des chercheurs membres du CIRANO afin de susciter échanges et commentaires. Ces cahiers sont rédigés dans le style des publications scientifiques et n'engagent que leurs auteurs.

The purpose of the Working Papers is to disseminate the results of research conducted by CIRANO research members in order to solicit exchanges and comments. These reports are written in the style of scientific publications. The ideas and opinions expressed in these documents are solely those of the authors.

Le CIRANO est un organisme sans but lucratif constitué en vertu de la Loi des compagnies du Québec. Le financement de son infrastructure et de ses activités de recherche provient des cotisations de ses organisations-membres, d'une subvention d'infrastructure du gouvernement du Québec, de même que des subventions et mandats obtenus par ses équipes de recherche.

CIRANO is a private non-profit organization incorporated under the Quebec Companies Act. Its infrastructure and research activities are funded through fees paid by member organizations, an infrastructure grant from the government of Quebec, and grants and research mandates obtained by its research teams.

Les partenaires du CIRANO – *CIRANO Partners*

Partenaires Corporatifs - Corporate Partners	**Partenaires gouvernementaux - Governmental partners**	**Partenaires universitaires - University Partners**
Autorité des marchés financiers	Ministère des Finances du Québec	École de technologie supérieure
Banque de développement du Canada	Ministère de l'Économie, de l'Innovation et de l'Énergie	École nationale d'administration publique
Banque du Canada	Innovation, Sciences et Développement Économique Canada	HEC Montréal
Banque Nationale du Canada	Ville de Montréal	Institut national de la recherche scientifique
Bell Canada		Polytechnique Montréal
BMO Groupe financier		Université Concordia
La Caisse		Université de Montréal
Énergir		Université de Sherbrooke
Hydro-Québec		Université du Québec
Intact Corporation Financière		Université du Québec à Montréal
Mouvement Desjardins		Université Laval
Power Corporation du Canada		Université McGill
Pratt & Whitney Canada		
VIA Rail Canada		

Le CIRANO collabore avec de nombreux centres et chaires de recherche universitaires dont on peut consulter la liste sur son site web. CIRANO collaborates with many centres and university research chairs; list available on its website.

© Octobre 2026. Vincent Boucher, Josette Gbeto, and Maripier Isabelle. Tous droits réservés. *All rights reserved*. Reproduction partielle permise avec citation du document source, incluant la notice ©. *Short sections may be quoted without explicit permission, if full credit, including © notice, is given to the source.*

Les idées et les opinions émises dans cette publication sont sous l'unique responsabilité des auteurs et ne représentent pas les positions du CIRANO ou de ses partenaires. *The observations and viewpoints expressed in this publication are the sole responsibility of the authors; they do not represent the positions of CIRANO or its partners.*

ISSN 2292-0838 (version en ligne)

Networks and Queues: Evidence from Medical Referrals in Canada [*]

Vincent Boucher[†], *Josette Gbeto*[‡], *Maripier Isabelle*[§]

Abstract

When prices are fixed, queues become central allocation mechanisms. Yet decentralized waiting queues are often unobserved, creating inefficiencies and the opportunity for informal access through referral networks. This paper studies a 2016 reform in Quebec, Canada, that introduced a centralized system for allocating medical referrals from general practitioners to physician specialists. We find that centralization significantly reduced new specialist consultations. We also infer physician referral networks and estimate a structural model that enables us to conclude that specialists tend to defer consultations for patients referred by general practitioners from whom they receive an unusually large number of referrals.

Résumé

Lorsque les prix sont fixes, les files d'attente deviennent des mécanismes centraux d'allocation des ressources. Toutefois, les files d'attente décentralisées sont souvent difficiles à observer, ce qui engendre des inefficacités et crée des possibilités d'accès informel par l'intermédiaire des réseaux de référence médicale. Cet article étudie une réforme mise en œuvre en 2016 au Québec, au Canada, qui a instauré un système centralisé d'attribution des demandes de consultation adressées par les médecins généralistes aux médecins spécialistes. Nous constatons que cette centralisation a entraîné une réduction statistiquement significative du nombre de nouvelles consultations auprès de médecins spécialistes. Nous reconstituons également les réseaux de référence entre médecins et estimons un modèle structurel qui nous permet de conclure que les médecins spécialistes ont tendance à reporter les consultations des patients orientés par des médecins généralistes dont ils reçoivent un nombre inhabituellement élevé de demandes de consultation.

[*] We thank the participants to the internal seminars on the Economics department at Université Laval, as well as those to the seminars at Georgetown University and Virginia Tech University. We acknowledge financial support from Mitacs (IT21158). Boucher thanks the SSHRC for financial support. Gbeto thanks the FRQSC for financial support. Isabelle is grateful for the support of the Sentinel North Research Chair in Economics and Brain Health. Ethics approval CERUL 2019-285. Data were obtain through the Commission d'accès à l'information du Québec through the Institut de la statistique du Québec. Ref RAMQ 7213-2021-10976 et CAI 1025507-S. This manuscript was reviewed using Refine.ink. All omissions and judgment calls are ours.

[†] Department of Economics, Université Laval, CRREP, CREATE; CIRANO; vincent.boucher@ecn.ulaval.ca

[‡] Department of Epidemiology and Biostatistics, Western University; jgbeto@uwo.ca

[§] Department of Economics, Université Laval, CRREP; mariepier.isabelle@ecn.ulaval.ca

Mots-clés/Keywords : Réseaux sociaux, files d'attente, références médicales / Social Networks, Queues, Medical Referrals.

Codes JEL/JEL Codes : I11, L14, J22

Pour citer ce document / To quote this document
Boucher, V., Gbeto, J., & Isabelle, M. (2026). Networks and Queues: Evidence from Medical Referrals in Canada (2026s-27, Cahiers scientifiques, CIRANO.)
https://doi.org/10.54932/WIMU8594

1 Introduction

In the absence of price adjustments, other mechanisms for allocating scarce resources necessarily emerge. In the context of publicly provided healthcare, waiting queues are widespread and determine when patients obtain access to physicians (Batabyal and Beladi, 2008). While prices are typically observable and summarize dispersed information about scarcity, physician-specific waiting lists are often decentralized, and only imperfectly observed by patients. This implies that patients cannot identify the shortest queue and results in highly heterogeneous waiting times, even among patients with similar needs.[1] Moreover, because patients cannot (in principle) move ahead in line, informal channels may become valuable: personal connections, physician discretion, or other forms of favoritism can substitute for explicit priority rules. (e.g., De Santana et al., 2018; Veinot et al., 2012; Zeltzer, 2020)

In this paper, we study the impact of centralizing the management of waiting queues in the healthcare system. We leverage the introduction of a 2016 reform aimed at centralizing the management of waiting queues for access to a subset of medical specialties in the province of Quebec, Canada. Prior to the reform, patients referred to specialized medicine by their general practitioner (GP) had to personally send their referral to a medical establishment or a specific specialist physician (SP), without much information about the length of waiting queues... unless they (or their GP) had private information or personal contacts. The *Centre de Répartition des Demandes de Services* (CRDS) implemented in 2016 are regional entities acting as a single point of contact to which GPs send their referrals, delegatin to a central entity the role of prioritizing them and reallocating them to physician specialists. The participation to this mechanism is voluntary for referring GPs and SPs, but encouraged by government entities.

Using administrative billing data from the *Régie de l'assurance maladie du Québec* (RAMQ), covering the period from 2011 to 2019, we evaluate the impact of this centralization on referral patterns, service volumes, and SPs' income. Our empirical strategy combines difference-in-differences, event-study approaches, and quantile treatment effects methods, enabling us to estimate both the average and distributional effects of the policy across physicians and patients.

The effect of a centralized system such as the CRDS is *a priori* unclear. On the one hand, centralized referral systems have been proposed as a way to improve coordination and equity in patient allocation. By providing a global view of referral flows and SPs availability, such systems can reduce inefficiencies and prevent certain SPs from being overburdened while others remain underutilized (Milakovic et al., 2021). On the other hand, centralization may also weaken the

[1] Waiting time represent the (time) cost of accessing care in a universal access system.

informational advantages GPs have about their patients' needs and trusted care providers(e.g., de Carvalho Nunes et al., 2016; Hackl et al., 2015; Brosig-Koch et al., 2022).

We show that, despite a relatively weak adoption, the introduction of the CRDS significantly reduced the average number of new patient consultations by SPs. While we provide some evidence that the system helped to redistribute referrals more evenly across available SPs, reducing the concentration of referrals to a small number of preferred physicians, the overall effect on the volume of new consultations is highly negative. We also observe an upward, though statistically insignificant, trend in the total number of medical services provided by SPs after the introduction of the CRDS. Notably, the volume of hospital-based services increased. We interpret these findings as evidence that the reform led SPs to reallocate their time away from new referrals towards their existing roster of patients, and not to a decrease in the overall medical services provided.

A particular feature of our setting is the reform took place in an environment characterized by long waiting queues. As such, we may expect strong dependence between outcomes (e.g. number of consultations) at different periods, which affects the interpretation of treatment effects and parallel trend assumptions. (Bonhomme, 2025; Cornwall and Wentland, 2025) We thus propose a structural model that studies the optimal choice of service provision for SPs. SPs dislike long queues of patients waiting, but may have heterogeneous preferences depending on whether patients were referred by a connected GP or not. In order to estimate the model, we first develop an estimator of the referral network between GPs and SPs that enables us to infer connections between providers. The estimator is based on the observed volume of referrals between pairs of GPs and SPs. We find that the referral network is sparse, with most SPs (78%) sharing too few referrals to be considered as having a connection. The average number of connections for SPs with at least one connection is 3.2, and the degree distribution has a fat tail and is approximately power-law distributed, as in many social and economic networks.[2]

We estimate our structural model using the estimated network structure. We find that the utility cost of a queue is *lower* for patients referred by a connected GP. Because connected GPs refer greater volumes of patients, we interpret these findings as evidence that SPs value equity of access for patients. Moreover, these preferences are sufficient to compensate for differences in referral volumes: the number of patients seen relative to the length of the queue is lower for patients referred by connected GPs.

Our contribution in this paper is twofold. First, we provide causal evidence on the impact

[2]See Jackson et al. (2017), and Figure A.1.

of centralizing medical referrals on physician behaviour by exploiting the introduction of the CRDS as a quasi-natural experiment. The medical literature on centralized (or single entry) systems for medical referrals suggests that centralized systems are associated with a reduction in waiting time (Milakovic et al., 2021). Di Renna et al. (2024) and Vo et al. (2022) find a similar association between the presence of centralized systems and lower waiting times in the Canadian context. However, some authors find that centralized systems weaken responsibility from patients and physicians and are less flexible, weakening the relationship between patients and their physicians. (Damani et al., 2017; Greenwood-Lee et al., 2018; Kennedy et al., 2025). Our findings contrast with the literature on waiting time, which may suggest unaccounted endogeneity issues in previous studies.[3] Our findings also highlight the fact that changing the allocation of referral requests does not only change the distribution of patients across SPs, but also the number of new consultations following a change in SPs' service-provision choices.

Second, and while we do not observe physicians' individual characteristics (e.g. gender or time/place of study), we show that the introduction of a voluntary centralized system strongly affects referral patterns. The literature on the impact of physicians' networks shows that physicians' gender has a strong impact on referral patterns and affect female physicians' income in particular (Zeltzer, 2020; Sarsons, 2017). Prior shared work or classroom environments have also been shown to lead to preferential referrals (Hackl et al., 2015). In complement to this literature, we find that despite biased referral patterns, SPs value equity among patients, and, everything else equal, prioritize patients referred by unconnected GP.

The remainder of the paper is structured as follows. In Section 2, we present the institutional context and describe the data. In Section 3, we outline the empirical strategy and reduced-form identification methods. In Section 4, we discuss the main empirical findings. In Section 5, we estimate the referral network, and we present and estimate the structural model in Section 6. Finally, we conclude in Section 7.

2 Context and data

In Quebec, a provincial jurisdiction within Canada, health care is provided at no monetary cost for patients.[4] Physicians' income and working conditions are jointly negotiated with the

[3] Another possibility is that the particular design or implementation of the CRDS was comparatively poor.

[4] The role of the private sector is very limited. For example, in 2016 (the time of the reform we exploit in this paper) the average household expenditure for medical services (excluding dental care and drugs) was 214 CAD. (Statistics Canada. Table 11-10-0222-01 Household spending, Canada, regions and provinces). Physicians choosing to see patients in the private sector must completely stop seeing patients under the public insurance regime, thus leading to a negligible fraction choosing this path.

provincial government so that prices (mostly fees for medical services, and wages in some specific settings) are largely unresponsive to changes in supply and demand. As a result, the health care system is characterized by long waiting queues across all types of medical services. This is especially true for access to specialized medicine.[5] In Quebec, patients cannot directly access most publicly provided specialized health care, and must first be referred by a general practitioner (GP).[6]

In this paper, we study the impact of a reform affecting the referral process between GPs and SPs. Before 2016 – when the reform was first introduced – patients who were referred for a consultation in specialized medicine were responsible for arranging their own appointment with the SP. In practice, this meant that patients had to send their referral to (usually many) medical establishments, and wait for a call back. While some patients were suggested to send their referral to a specific medical establishment or specialist by their GP (some GP cabinets even sending the referral requests on behalf of their patients), many were not. The efficiency of the referral process was therefore highly heterogeneous, depending on the level of familiarity of patients (and that of their GP) with the local healthcare system, including the different SPs accepting new patients and their respective waitlists. This inefficient organization of medical services negatively affected both patients and healthcare professionals (Clair, 2000). The lack of coordination between healthcare providers required patients to navigate their own medical journey. These organizational shortcomings, along with an uneven workforce distribution and various incentives embedded in physicians' remuneration schemes ultimately led to what governments considered unreasonable waiting times to access specialized medicine. The creation of *Centres de répartition des demandes de services* (CRDS), aimed at improving access to SPs.

2.1 The reform

In 2016, in an effort to increase efficiency and reduce waiting lists, the province introduced a centralized referral system between GPs and SPs for a subset of specialties (the treated group). The CRDS is a single entry point for GP in primary care who wish to refer their patients for an initial consultation with a SP. The structure was simultaneously implemented in all regions of the province, even if each regional structure is run independently.

The CRDS acts as an intermediary between GPs and SPs. For each new referral going

[5] See for example https://www150.statcan.gc.ca/n1/daily-quotidien/250729/dq250729a-eng.html for a brief overview of waiting times to see a physician specialist in Canada.

[6] An exception are referrals from emergency room physicians. These referrals are excluded from our analysis, see Section 2.2 below.

through the CRDS, GPs need to send a referral request to the regional management system, which then has the responsibility to find an SP for the patient. Submitting a request to the CRDS is a very structured process. In particular, the GP has to specify the reason for the referral and to identify a clinical priority for the patient. If there is no option in the system that corresponds to the patient's clinical condition, the GP must specify an alternative reason for consultation and add a clinical level of priority (from A to E, the latter denoting the least urgent cases).[7]

One notable feature of the CRDS is that physicians are not mandated to use it. GPs can still refer their patients through their personal network–or specify a particular SP in requests sent to the CRDS (known as referral by name). As a result, GPs may choose the referral method based on several factors, including the urgency of the case, geographic location, expected wait times, local norms or administrative guidelines, and trust in a specific colleague.[8] While there is limited official information on the overall adoption of the CRDS, we estimate that it is relatively low, and heterogeneous across specialties.[9] In order to allocate patients efficiently, a regional branch of the CRDS must be aware of the services offered by SPs in their region. Participating SPs are asked to submit a list of service offered (e.g. a subset of reasons for consultations for which they have an expertise), the type and location of the medical establishment they are affiliated to, and their available time slots for predetermined periods of time.[10]

The CRDS was implemented in three phases (Fall 2016, Fall 2018 and Fall 2019), each phase gradually including more specialties. We leverage this feature of the implementation in our reduced-form empirical strategy, comparing *CRDS-treated* specialties with not-yet treated ones [11] We focus only on the initial implementation phase, using specialties treated in Fall 2019 as our control group, and discarding the data for the specialties treated in Fall 2018.[12] The specialties

[7]The rules by which the CRDS operate are generally similar across regions (and in particular those covered in our sample). For example, it is recommended that general practitioners contact an on-call SP directly if their patient's clinical priority is A (indicating that the patient is in critical need of care and should be seen within three days or less). However, there are some differences in the referral process across specialties and across regions. For example, orthopedic SPs should not receive referrals for priority A or B through the CRDS without the GP first consulting with an on-call spinal neurosurgeon. In urology, certain CRDS (for example in the Quebec city region) refuse to process requests for vasectomies because the service is offered in private clinics.

[8]Although GPs can continue to refer patients to the same SP by naming them in the CRDS request, they are no longer guaranteed that the patient will be seen by the designated SP. This uncertainty may reduce any implicit pressure on the SP stemming from personal or professional networks.

[9]See Appendix A for more details. In particular, we estimate that the adoption rate ranges from 10% to 45% for the treated specialties in our sample, measured as the fraction of consultations for which the referral went through the CRDS.

[10]There is some heterogeneity across regions with respect to how long in advance SPs have to submit their availability. For example, in the Chaudière-Appalaches region, it is recommended for SP physicians to provide their clinic schedules two months in advance, while in the Montreal region, the availability form completed by the SP must be sent to the CRDS at least two weeks before the first date of availability registered.

[11]A list of treated specialties for each phase is available in Appendix A.

[12]Because of the COVID 19 pandemic, which significantly affected the medical system for a prolonged period,

were allocated to an implementation phase mostly based on the specialty-level historical average volume of service, correlating with waiting time for certain procedures, generally subject to high media coverage.

2.2 Data and Descriptive Statistics

We use administrative data on physicians' billing from the *Régie de l'assurance maladie du Québec* (RAMQ), which records all fee-for-service medical activities provided by health professionals for treated and untreated specialties practicing in regions of Quebec between April 2010 and September 2020.[13] The data include detailed information on the physician performing the medical service, their specialty, the invoice number for each medical service, the date of service, the type of medical service, the amount billed, the identifier and specialty (including general practice) of the referring healthcare professional for the service billed[14], as well as the type of facility and locality in which the medical service is provided. We note that fee-for-service is the main form of payment for physicians in Québec, totaling more than 77% of all clinical payments made to physician by the end of our sample period in 2020 (for Health Information", 2022), and an overwhelming proportions of physicians' activities which result from referrals.

We restrict our sample to medical services provided by specialists (SPs). As mentioned above, we define treatment as the first implementation phase (Fall 2016) of the CRDS. Treated specialties are Cardiology, Urology, Gastroenterology, Otolaryngology (ORL), and Pediatrics. Specialties forming our control group are those unaffected by the implementation of the CRDS within our sample period: Plastic Surgery, Dermatology, and Pneumology. We exclude from our sample all physician in specialties that entered the CRDS in the second implementation phase of 2018, focusing on a simple difference-in-differences design rather than on a staggered one given that our post-period is limited to the pre Covid-19 months. We also exclude physicians who have not billed any services for at least four consecutive quarters, as these are likely retired or on-extended-leave practitioners.

We aggregate data at the annual level to account for potential seasonal variations in patient flows and referral patterns across specialties. For example, general practitioners may refer more patients to pneumonologists during winter months for pneumonia and chronic lung disease exacerbations, while referrals to gastroenterologists may increase in summer for gastrointesti-

we do not use 2020 data in our analyses. We therefore shy away from opting for an empirical design leveraging the staggered implementation of the CRDS.

[13] List of specialties and regions can be found in Appendix A.

[14] A referring physician identifier appears only for a first contact between a patient and the physicians they have been referred to

nal conditions. Annual aggregation mitigates short-term fluctuations and measurement error, providing a more stable basis for comparing treatment and control groups across time.

Table 1 summarizes the main characteristics of our sample over the pre-treatment period (2011–2015), separately for treated and untreated specialties. Panel A documents the scale of the dataset[15] The treated group comprises 1,420 specialists and the control group 450, corresponding to 12,790 and 4,030 SP-year observations, respectively. This sample covers approximately 1.35 million GP–SP pairs observed in the treated group and 0.57 million in the control group, generating respectively 12.1 million and 5.1 million GP–SP pair × year observations. The number of GPs referring to treated and control specialists is 30,465 and 27,240 (with overlap), respectively, suggesting broad coverage of Quebec's primary care network across both groups.[16]

Panel B reports specialist-level outcomes. Several differences in pre-treatment levels are observed between treated and control specialties. Treated specialists receive fewer new GP referrals per year than control specialists (545.8 vs. 653.4, $p = 0.010$) and have fewer total invoices (2,165.6 vs. 2,449.0, $p = 0.040$). Income from new referrals is also lower for treated specialists, although the difference is only marginally significant (CAD,55.3K vs. 61.9K, $p = 0.069$). In contrast, total medical services per SP are not statistically different across groups (3,212.1 vs. 3,408.5, $p = 0.306$), and total income per SP is also statistically similar (CAD,222.4K vs. 205.7K, $p = 0.107$). The absence of a statistically significant difference in average total income is reassuring, since it suggests that treated and control specialists were not on sharply different economic levels before the reform.[17]

These differences in pre-treatment levels likely reflect differences in specialty mix, referral-network structure, and the smaller and more concentrated composition of the control group. Table 2 further documents substantial heterogeneity across specialties within each group. For example, ORL specialists receive on average 1,143 new referrals per year, compared to only 195 for pediatricians, reflecting differences in clinical scope rather than differential pre-reform dynamics. Finally, a slightly larger share of treated specialists practice in hospital settings relative to control specialists (98.3% vs. 96.0%, $p = 0.020$). Although statistically significant, the magnitude of this difference is small, and both groups are overwhelmingly hospital-based.

Panel C characterizes the GP–SP pair-level structure of the data. The referral network is

[15] For confidentiality reasons, sample sizes and counts are rounded to the nearest multiple of five, following the disclosure rules of the Institut de la statistique du Québec, through which we access the data.

[16] Note that the same GP may refer patients to both treated and untreated specialists; the counts in Panel A therefore overlap across groups.

[17] There is however well documented heterogeneity of incomes within treated and untreated specialties, even if average incomes are similar, see e.g. Canadian Institute for Health Information. https://www.cihi.ca/en/indicators/average-gross-clinical-payment-per-physician Accessed September 22, 2026.

highly sparse: only 19.1% of treated pair-years and 17.9% of control pair-years record at least one consultation, although this difference is statistically significant ($p = 0.031$). Conditional on having at least one consultation, treated pairs average 2.85 consultations per year, compared with 2.71 for control pairs ($p = 0.041$). The unconditional number of consultations per pair is also slightly higher among treated specialties (0.544 vs. 0.485, $p = 0.007$). Nevertheless, the distribution of positive consultations remains very similar across most of the distribution: the 25th, 50th, and 75th percentiles are identical across groups. The main difference appears around the upper tail, with the 90th percentile equal to 5 among treated pairs and 6 among control pairs, although the 99th percentile is 24 consultations in both groups.

Treated specialists also receive referrals from fewer distinct GPs on average than control specialists (244.8 vs. 322.7, $p < 0.001$). This suggests that pre-reform referral networks were somewhat more concentrated among treated specialties, with referral relationships involving fewer referring GPs but slightly higher pair-level intensity. This pattern is consistent with the motivation for the CRDS, which aimed to reduce dependence on established bilateral GP–SP referral relationships and to redistribute referrals through a centralized mechanism.

Overall, the pre-treatment comparisons in Table 1 show that treated and control specialties differ in some baseline levels. These differences are expected given the smaller size of the control group and the heterogeneity in clinical scope across specialties. Importantly, such level differences do not by themselves invalidate the research design, since the identification strategy relies on the absence of differential pre-reform trends rather than equality of pre-treatment means. We therefore complement these descriptive comparisons with event-study graphs in Section 3, which provide direct evidence on the parallel-trends assumption.

3 Reduced-form analysis

To first identify the causal impact of referral centralization on physician networks and SP outcomes, we employ a difference-in-differences (DiD) and event-study framework, leveraging variation in the timing and scope of CRDS implementation across medical specialties.

Our baseline specification is as follows:

$$Y_{ij,t} = \sum_{s \neq 2015} \mathbf{1}[s = t] \times \beta_s \times D_j + \lambda_{ij} + \gamma_t + \varepsilon_{ij,t} \tag{1}$$

where $Y_{ij,t}$ is the outcome of interest. For referral volumes, it refers to the number consultations for new referrals between GP i and SP j; for other outcomes, such as SP-level activity or

	Treated	Untreated	Difference	p-value
Panel A: Sample Composition				
Number of specialists	1,420	450	970	
Number of GPs (referring)	30,465	27,240	3,225	
Number of GP–SP pairs observed	1,346,610	568,800	777,810	
SP × year observations	12,790	4,030	8,760	
GP–SP pair × year observations	12,119,490	5,119,220	7,000,270	
Panel B: Specialist-Level Outcomes (annual averages, 2011–2015)				
New GP referrals per SP	545.83	653.40	-107.57**	0.010
	(661.44)	(841.79)	(41.64)	
Total medical services per SP	3,212.10	3,408.48	-196.37	0.306
	(3,156.70)	(3,827.60)	(191.92)	
Total invoices per SP	2,165.56	2,449.04	-283.48**	0.040
	(2,190.24)	(2,761.99)	(137.77)	
Income from new referrals per SP (CAD 000)	55.29	61.86	-6.57*	0.069
	(62.32)	(72.62)	(3.61)	
Total income per SP (CAD 000)	222.43	205.68	16.74	0.107
	(212.46)	(200.45)	(10.38)	
Share practicing in hospital (%)	98.3	96.0	2.4**	0.020
	(12.8)	(19.7)	(1.0)	
Panel C: GP–SP Pair-Level Outcomes (annual averages, 2011–2015)				
Share of pair-years with ≥ 1 consultation (%)	19.1	17.9	1.2**	0.031
	(39.3)	(38.3)	(0.6)	
Annual consultations per pair (if > 0)	2.85	2.71	0.14**	0.041
	(6.07)	(7.23)	(0.07)	
Annual consultations per pair (unconditional)	0.544	0.485	0.060***	0.007
	(2.881)	(3.227)	(0.022)	
Percentiles of consultations per pair (if > 0):				
p25	1	1		
p50	1	1		
p75	2	2		
p90	5	6		
p99	24	24		
Avg. GPs referring to a given SP (per year)	244.79	322.70	-77.91***	< 0.001
	(185.51)	(238.55)	(13.70)	

Notes: The pre-treatment period covers fiscal years 2011–2015. Treated specialties are Cardiology, Urology, Gastroenterology, ORL, and Pediatrics. Untreated specialties are Plastic Surgery, Dermatology, and Pneumology. Differences are computed as treated minus untreated. Standard deviations are reported in parentheses under the treated and untreated means; standard errors are reported in parentheses under the difference column. Income variables are deflated to 2015 CAD. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
The treated and untreated groups differ in some pre-treatment levels. These differences may partly reflect the smaller and more concentrated composition of the untreated group, as well as underlying differences in specialty mix and referral-network structure. In particular, treated specialties have fewer new GP referrals per SP and fewer invoices, while income from new referrals is marginally lower. However, total medical services and total income per SP are not statistically different across groups. These baseline differences are therefore interpreted as differences in levels rather than evidence of differential pre-reform trends; the event-study graphs in Section 4 support the parallel-trends assumption.

Table 1: Descriptive Statistics: Pre-Treatment Period (2011–2015)

Specialty	Group	N specialists	New referrals per SP	Total services per SP	Total income per SP (CAD 000)
Treated Specialties					
Cardiology	Treated	383	555.0	3,602.9	274.7
Gastroenterology	Treated	202	770.4	3,472.1	249.2
ORL	Treated	181	1,142.8	5,191.6	257.9
Pediatrics	Treated	534	194.7	1,968.7	150.0
Urology	Treated	121	798.4	4,067.7	278.8
Total treated		**1,421**			
Control Specialties					
Plastic Surgery	Control	89	492.5	2,494.5	225.1
Dermatology	Control	140	1,011.7	4,730.9	192.0
Pneumology	Control	219	489.7	2,934.5	206.6
Total control		**448**			

Notes: All variables are annual averages over the pre-treatment period 2011–2015. Treated specialties were affected by the CRDS reform in Fall 2016. Control specialties were scheduled for treatment after Fall 2019. Income is deflated to 2015 CAD.

Table 2: Specialty-Level Descriptive Statistics: Pre-Treatment Period (2011–2015)

income, it is indexed by SP j alone. The variable D_j is a treatment indicator equal to 1 if the observation corresponds to the treatment group (i.e. if SP j is from a treated specialty), and 0 otherwise.

The parameter of interest, β_s, captures the average *intention-to-treat* effect, since we do not observe whether the SP actually uses the CRDS. We only observe that their specialty is treated (i.e. the CRDS platform is deployed and functional for their specialty). The specification also includes individual fixed effects λ_{ij} (or λ_j) and time fixed effects γ_t. We exclude the year prior to treatment to avoid perfect collinearity in the event-study framework. Standard errors are clustered at the GP–SP pair level for referral outcomes, and at the SP level for SP-level outcomes.

3.1 Identification challenges

Identification notably relies on the parallel trends assumption: in the absence of treatment, treated and control groups would have followed similar trends in outcomes. Recent contributions to the DiD literature highlight that violations of this assumption can bias estimates, particularly in two-way fixed effects (TWFE) frameworks (e.g., Roth, 2022; Roth et al., 2023; Rambachan and Roth, 2023). To assess the robustness of our estimates, we implement the sensitivity analysis framework proposed by Rambachan and Roth (2023). The sensitivity analysis is presented in

Appendix B.3.[18] We find that our results are robust to violations of the parallel trend up to 1.5 times of the maximum pre-treatment violation, depending on the specification. However, our results are not directly robust to the inclusion of non-linear differential trends. This may not be surprising given the recent work by Cornwall and Wentland (2025) (see also Bonhomme (2025)) who make the point that the presence of autocorrelation may affect identification, and the statement of the parallel trend in particular. Indeed, as discussed in Section 2, the health care system in Quebec is characterized by long queues of patients waiting for a consultation with a SP. We therefore expect that any shock (observed or not) affecting the referral of patients would potentially affect the number of consultations over multiple periods (see Section 6). To address this, we also estimate the impact of the CRDS on the first-differences of our outcome variables. This strategy helps capture year-over-year changes and dynamic responses that may be obscured in level comparisons. (See Appendix B.2.) We also present difference-in-differences estimators allowing for linear trends and autoregressive terms, as suggested by Cornwall and Wentland (2025) (see Appendix B.1). Results are broadly consistent.

Another threat to identification is the concern that the selection of specialties included in the 2016 reform may not have been random. However, based on discussions with policymakers involved in CRDS implementation, we confirmed that the selection was based exclusively on the specialty-level historical average volume of service. Indeed, before the implementation of the centralized system, the policy makers had little knowledge of the number of referrals produced or on the number of patients waiting for a consultation. The information they had was aggregated, and mostly time-invariant. Thus, any impact of the choice of treated specialties on the treatment effect should be captured by the fixed effects, making the reform plausibly exogenous to physicians' referral behavior.

4 Results

We present our main reduced-form results by plotting the event-study D-i-D coefficients from equation 1. Our main takeaway is that the introduction of the CRDS led to reduced volumes of new consultations from SPs, but did not reduce to overall volume of services they provide. We also find that the treatment effect increases with time. This is intuitive since the system is characterized by long queues so that any permanent treatment is expected to accumulate through time. We present a formal analysis of this, using a structural model, in Section 6. In

[18] Appendix B.3 also presents placebo tests for alternative treatment date.

Appendix B.1, we present difference-in-differences estimates with and without linear trends, as well as with an autoregressive term, as suggested by Cornwall and Wentland (2025).

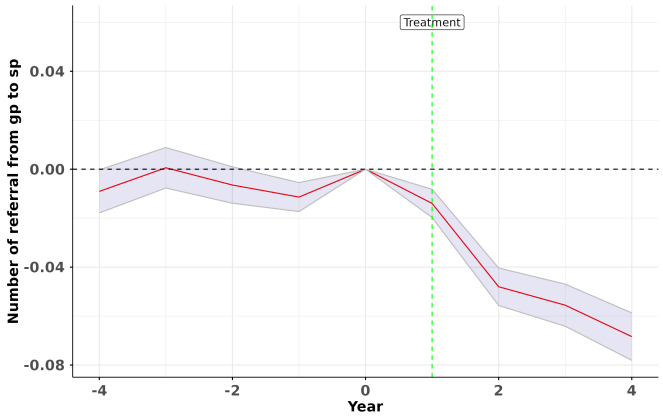


Figure 1: Patient referrals between GP–SP pairs

Figure 1 first illustrates the evolution of patient referrals between GP–SP *pairs*. Each point connected by the red line represents the estimated difference between treated and untreated specialties relative to the pre-reform year (2016). The shaded area displays the 95% confidence intervals derived from Equation (1).

Following the introduction of the CRDS, we observe an important and persistent decline in the number of new pairwise consultations for treated specialties. In the first year after implementation, treated specialties experienced approximately 0.039 fewer new consultations per GP–SP pair, a 7% decrease from pre CRDS levels. This gap widened in subsequent years, reaching about 0.042 (8%) in 2018 and 0.056 (10%) in 2019. These results imply that the introduction of the CRDS reduced, on average, the expected number of referrals for a given GP–SP pair. The same effect is found at the SP level, although confidence intervals are larger.[19] Figure 2 shows that treated SPs provided on average fewer new consultations following the introduction of the CRDS.

We then turn to the average volume of services provided by SPs, which includes consultation to newly referred patients, and all other billable medical services. Figure 5 shows that the introduction of the CRDS did not decrease the total volume of service provided. We interpret this finding as evidence of a reallocation of time previously spent on new consultations toward time spent providing services to a SPs' existing roster of patients. While this could potentially indicate

[19]This is partly mechanical as the observation units in Figure 1 are pairs ($nT = 24,762,740$) while they are SPs ($nT = 20,889$) in Figure 2.

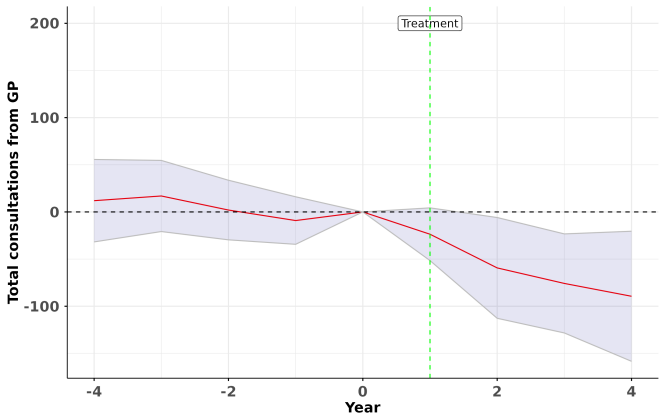


Figure 2: Volume of new referrals seen by SPs

a possible improvement in service intensity or continuity for these existing patients, it could also signal a moderation of the incentives under fee-for-service schemes, where overprovision may occur. In any case, these results are not aligned with the intended objective behind the implementation of the CRDS structure to facilitate access to physician specialists for patients needing to be referred.

This interpretation is further supported by Figures 3–5. The introduction of the CRDS decreased the income generated from new consulations, (Figure 3) but did not reduce the total income received by SPs (Figures 4 and 5)

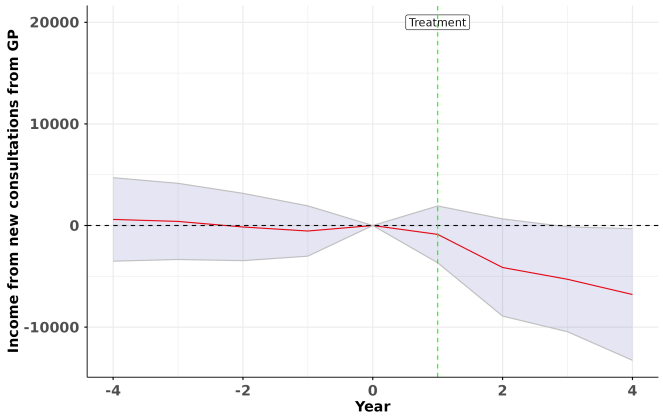


Figure 3: Income from new referrals from GP to SP

When using the method proposed by Callaway and Li (2019) to investigate distributional effects, we find that the treatment effect appears to be stronger for SPs in lower and higher

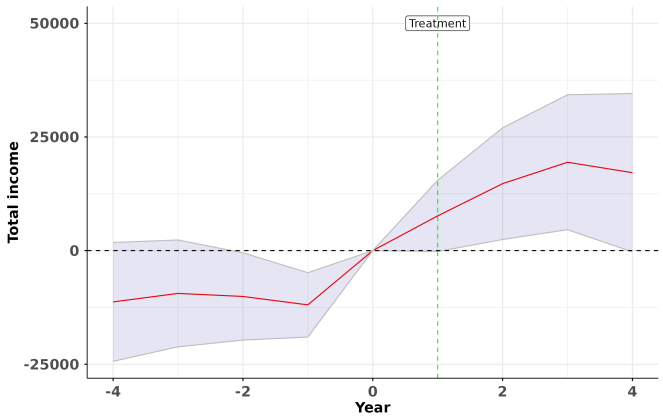


Figure 4: Income from total services provided by SP

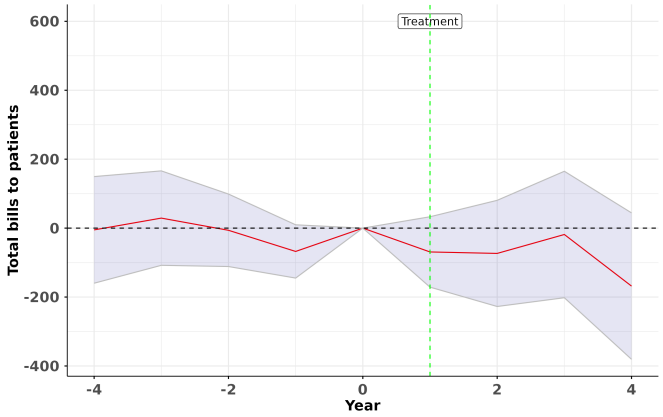


Figure 5: Total invoice volume billed by SP

outcome quantiles (Appendix Figure B.4). The results presented in this section roughly hold when we consider outcomes in first differences (Appendix Figure B.1), when we restrict our attention to the main type of establishment in with a SP offers his services (Appendix Figures B.5, B.6 and B.7) or to specific regions (Appendix Figures B.8, B.9 and B.10), albeit often suffering from an important loss of power. Our results also survive a placebo test in which we solely focus on pre-2016 data and estimate our model with three different fictitious reform dates (Appendix Table B.2 and B.3), and using the Honest D-i-D method proposed by Rambachan and Roth (2023) we find that our result remain significant in the presence of a violation in parallel trends that would reach 1.5 times the magnitude of the maximum pre-treatment violation, depending on the specification (Appendix Figure B.2). These robustness checks and sensitivity

analyses are presented in more detail in Appendix B.4.

5 Referrals' Network

In the previous section, we found that the introduction of the CRDS decreased the volume of new consultations for each *pair* of GP and SP. To put these results into context, we then aim to infer the identity of these pairs —given that some of them may share a special relationship (i.e. a link)—, which we will then use to estimate a structural model providing insights on the expected longer-run dynamics of the effects documented above, and on the mechanisms behind physicians' response to the implementation of a centralized referral mechanism.

Specifically, we propose a strategy that allows to recover the referral network (the set of links) from the data. Consider the referring GP i and SP j. Let $\mathbb{E}y_{ij}$ be the expected number of consultations emanating from referrals from i to j over some fixed period (say, annually). We denote $g_{ij} = 1$ if GP i and SP j are *linked*, and $g_{ij} = 0$ otherwise. Our identification assumption is the following:

Assumption 1. *There exists a function ψ such that $\mathbb{E}y_{ij} = \psi(j, g_{ij})$ for all referring physician i and SP j. The function ψ is such that $\psi(j,1) > \psi(j,0)$ for all j.*

Assumption 1 states that the identity of the GP does not affect the expected number of referrals received, conditional on the linking status with the SP. The expected number of referrals is however allowed to freely depend on the identity of the SP. Assumption 1 implies that, for any given SP j, the only source of heterogeneity in the long-term average number of referrals received from GPs is captured by the network.

We will use Assumption 1 in order to infer the network structure from the data. In some sense, this *defines* what we mean when we say that a GP and a SP are *linked*. Linked physicians do not necessarily know each others personally. Linked physicians are the ones for which there is a larger expected number of referrals.

Under Assumption 1, $\mathbb{E}y_{ij}$ can take at most two values, for each SP j: y_j^0 (if i and j are not linked), and $y_j^1 > y_j^0$ (if i and j are linked). Consider any SP $j : 0 < n_j < n$, where $n_j = \sum_i g_{ij}$ is the number of links that SP j has. Then, if we know any number θ_j such that $\theta_j \in (y_j^0, y_j^1)$, we can reconstruct the network structure, i.e. letting $g_{ij} = \mathbb{1}\{\mathbb{E}y_{ij} \geq \theta_j\}$.

Of course, in practice, we do not directly observe $\mathbb{E}y_{ij}$, and we do not know the set of SPs j such that $j : 0 < n_j < n$. We therefore propose the following estimation strategy. For each combination of specialties and regions, we assume that $\max_j y_j^0 < \min_{j:n_j>0} y_j^1$. As such, any

$\theta \in [\max_j y_j^0, \min_{j:n_j>0} y_j^1]$ allows to reconstruct the network structure. We will refer to this set as the *identified set.* We now present our estimator.

Let $\bar{y}_{ij}$ be the empirical average number of referrals from i to j, and let $y_j^- = \min_i \bar{y}_{ij}$ and $y_j^+ = \max_i \bar{y}_{ij}$. Define the following objective function:[20]

$$Q(\theta; \bar{\mathbf{y}}) = \sum_j \sum_i (\bar{y}_{ij} - y_j^-)^2 \mathbb{1}(\theta \geq \bar{y}_{ij}) + (y_j^+ - \bar{y}_{ij})^2 \mathbb{1}(\bar{y}_{ij} \geq \theta).$$

For a fixed value of θ, the function $Q(\theta; \bar{\mathbf{y}})$ gives the sum of the dispersions (sums of squares) below and above θ. Recall that θ is interpreted as a threshold allowing to reconstruct the network structure. When the average number of referrals between i and j is larger than θ, we infer that the pair ij is linked. If the long-term averages converge to the expected values, Assumption 1 implies that the dispersion below and above θ is null, for any value of θ inside the identified set.

While $Q(\theta, \bar{\mathbf{y}})$ can be used to recover the identified set, our objective is to recover the network structure and not the entire identified set. In practice, it is therefore more convenient to focus on a smoothed version of objective function:

$$Q^h(\theta, \bar{\mathbf{y}}) = \sum_j \sum_i (\bar{y}_{ij} - y_j^-)^2 F([\theta - \bar{y}_{ij}]/h) + (y_j^+ - \bar{y}_{ij})^2 F([\bar{y}_{ij} - \theta]/h))$$

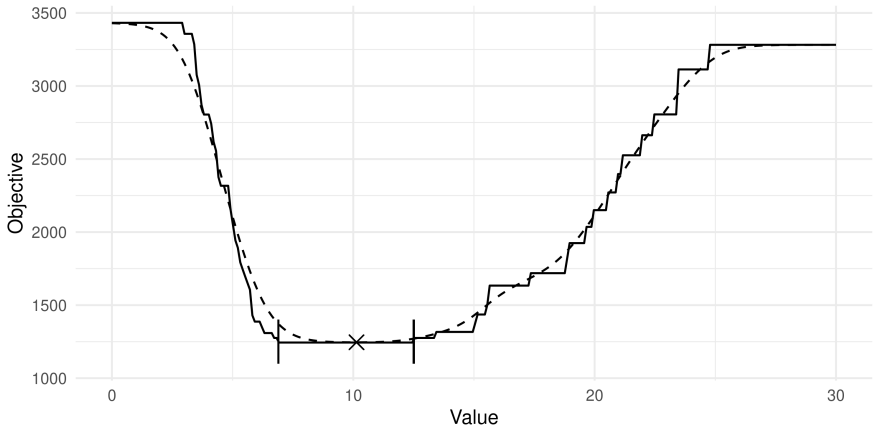
where F is a smooth distribution function with density f and where $h > 0$ is a small number. We will also assume that the density has full support, is symmetric around zero, has convex tails and is such that $lim_{x\to\infty} f(x) = 0$. This is similar to typical smoothing procedures (e.g. Kernel regressions) although here, h is fixed. In particular, we do not require that h goes to zero as the sample size grows.

An example of the distinctions between $Q(\theta, \bar{\mathbf{y}})$ and $Q^h(\theta, \bar{\mathbf{y}})$ is displayed in Figure 6 below. The next proposition shows that the estimator based on the smoothed objective function allows to consistently estimate a pseudo-true parameter which lies strictly inside the identified set.

Proposition 1. *Let* $\hat{\theta}^h = \arg\min_\theta \; Q^h(\theta, \bar{\mathbf{y}})$ *and suppose that a WLLN applies for* $\bar{\mathbf{y}}$*. Suppose also that there exists* j *such that* $0 < n_j < n$*, that Assumption 1 holds, and that* $\max_j y_j^0 < \min_{j:n_j>0} y_j^1$*. Then, there exists* h *small enough such that plim* $\hat{\theta}^h \in (\max_j y_j^0, \min_{j:n_j>0} y_j^1)$.

Proposition 1 shows that the estimator is (asymptotically) strictly inside the identified sets. Of course, consistency requires that the long-term average number of referrals converges in

[20] As mentioned above, the analysis is performed separately for each combination of specialty and region.



Note: Synthetic example of the empirical objective and smoothed objective functions for 60 different values of $\bar{y}_{ij}$. The set of values of θ that minimize the (not smoothed) empirical objective function $Q(\theta, \bar{\mathbf{y}})$ lies between the two short vertical lines. The estimated value for the smoothed empirical objective function $Q^h(\theta, \bar{\mathbf{y}})$ is displayed with a $\times$ sign.

Figure 6: Objective and Smoothed objective functions

probability to its expected value. While this is arguably not a primitive assumption, we present sufficient conditions based on a structural model in the next section.

The estimated thresholds (for each combination of specialty and region) are sufficient to define an estimator of the network: $\hat{g}_{ij} = \mathbb{1}\{\bar{y}_{ij} \geq \hat{\theta}_i\}$. Moreover, because $plim\ \hat{\theta}^h$ is in the interior of the identified set, it implies that the estimating uncertainty around $\hat{\theta}^h$ can be ignored. Indeed, this is due to the fact that $\hat{g}_{ij}$ is constant in a neighborhood around $plim\ \hat{\theta}^h$.

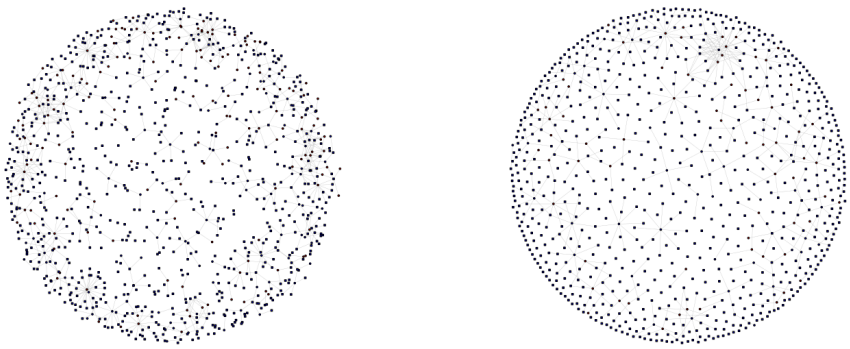
We estimate θ separately for each specialty and region, as these are two factors that can be strongly associated with referral volumes. Results are presented in Table 3. Overall, we see no systematic difference between (pre-treatment) treated and untreated specialties. Region 1 (including Montréal, the province's largest city and economic centre, and the surrounding regions of Laval, Montérégie, Lanaudière and Laurentides) has generally larger threshold values, which is intuitive since it is the most populous in our sample.

Using estimated thresholds in Table 3, we can construct the referral network. Figure 7 presents a visualization of the network and Figure 8 the implied degree distribution. Pre-treatment, there is a larger fraction of SPs with no link in treated specialties, relative to treated specialties. Figure A.1 of the Appendix provides evidence that the referral network has a fat tail and is roughly power-law distributed.

Specialty	06*	09	11	12*	17	25*	27*	31*
Region 1: Montréal, Laval, Lanaudière, Laurentides, Montérégie								
	47.907	23.615	68.658	40.701	305.719	81.324	75.640	31.911
Region 2: Québec, Mauricie-Centre-du-Québec, Chaudière-Appalaches								
	35.569	7.066	37.856	46.560	14.548	34.898	8.978	16.114
Region 3: Estrie								
	12.901	11.288	24.591	20.755	3.172	47.649	4.164	41.252

Note: Each cell corresponds to a different estimation using the smoothed objective function $Q^h(\theta, \mathbf{y})$ on *pre-treatment* data. The threshold values are in number of referrals per year between a given pair of physicians. Region 1. Standard errors are omitted since estimated parameters asymptotically fall strictly inside the identified set for the un-smoothed objective function. A star (*) next to a specialty code signifies that the specialty is (eventually) treated.

Table 3: Estimated threshold values

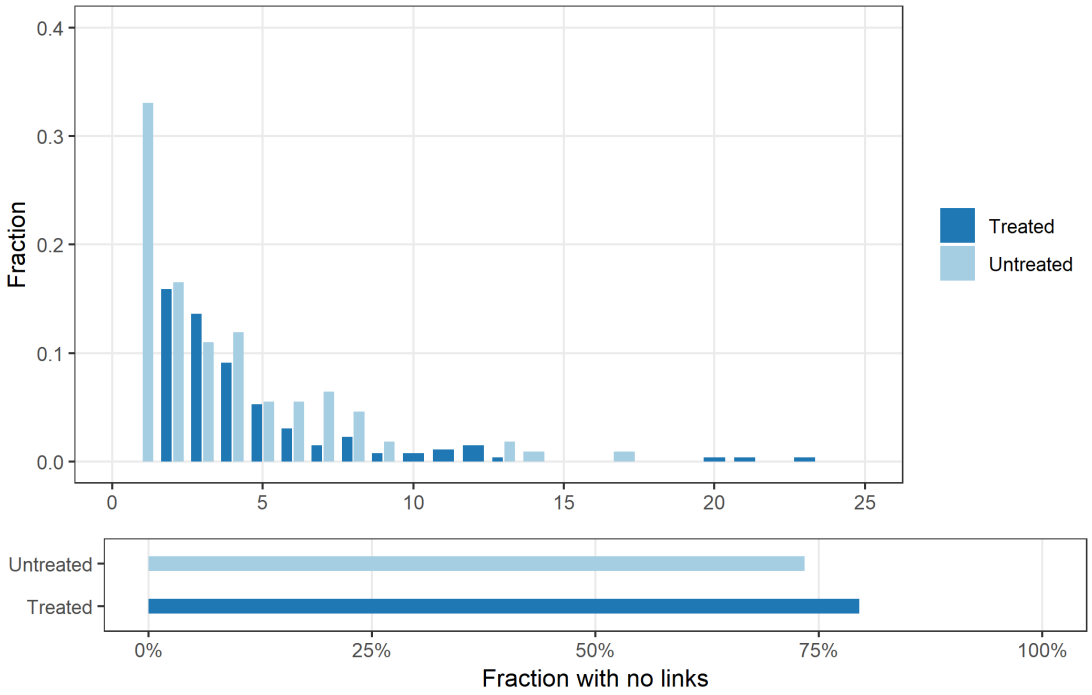


(a) Referral network (treated specialties) (b) Referral network (untreated specialties)

Figure 7: Referral network

6 Structural Model of referrals

The objective of the structural model we present here is twofold. First, it provides insights on the results from the event-study graphs presented in section 4. In particular, we can show formally that, under reasonable assumptions, the magnitude of the treatment effect we estimate should increase over time. As discussed previously, this is due to the fact that there is excess demand at each period, which results in the formation of queues. Our model thus enables us to better understand changes in the slopes of our event-study graphs post treatment, and to think about the longer-run dynamics of the reform we study even when our sample period –because of the



Note: Top graph presents an histogram for the degree distribution, conditional on having at least one link. The bottom graph presents an histogram for the fraction of SPs with no link.

Figure 8: Degree distribution for the referral network

eclosion of the Covid-19 pandemic and its impact on the healthcare system– doesn't. Second, the model can be estimated with the data and allows to answer additional questions that enrich our interpretation of the reduced-form results. In particular, we show that the utility cost for SPs of having long queues of patients waiting for access is lower when patients are referred by a *linked* GP. Given how the network structure is inferred, this implies that the disutility for SPs of having long queues of patients waiting is smaller when those patients are referred by a GP who, on average, sends them an unusually large number of referrals.

The model aims to capture the main features of the referral process pre-CRDS. Although it could be adapted to analyze post-CRDS referrals, its analysis is complicated by the short post-treatment period (partly due to the Covid-19 pandemic) and the fact that our data does not enable us to observe which referral is sent through the CRDS, and which one is not. We therefore present and estimate the model for the *pre-treatment period only*.

We first suppose there is a population of m SPs and n general practitioners (GPs). SPs and GPs are linked through a (bipartite) network. Recall that $g_{ij} = 1$ if GP i is linked to SP j and

$g_{ij} = 0$ otherwise. Recall also that n_j denotes the number of links that SP j has.

The model evolves in discrete time and mainly focuses on the service-provision choices by SPs. At period t, SP j observes $a_{ij,t}$, the queue of patients referred by i before t (but not yet seen by j) and chooses to see $c_{ij,t} \geq 0$ of these patients.[21] SP j can also see patients from their own roster, $c_{0j,t}$. At the end of period t, GP i refers $b_{ij,t} \geq 0$ new patients to SP j. Unseen patients at time t are waiting (queuing) until SP j sees them.

This dynamic process can be expressed compactly using matrix notation as follows.

$$[\mathbf{A}_{t+1} - \mathbf{A}_t] = \mathbf{B}_t - \mathbf{C}_t, \tag{2}$$

where the matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are $n \times m$ non-negative matrices. The variation in the length of the queues between t and $t+1$ (left-hand side) are given by the difference between new referrals, and consultations at time t (right-hand side).

The matrix of referrals (i.e. $\mathbf{B}$) partly reflect GPs' choices, but also that of the patients. In the absence of a centralized system a GP can either provide the patient with a referral for "any" SP of a given specialty, or encourage them to see a SP in particular. We therefore expect $\mathbf{B}$ to be a function of the network structure. The matrix of consultations (i.e. $\mathbf{C}$) is the result of SPs' choices. Matrix $\mathbf{C}$ is *the only observed matrix* in our data. We assume that it is the result of SPs optimal service-provision choices.

Specifically, we assume that at each period t, SP j chooses the vector $(\mathbf{c}_{j,t}, c_{0j,t})$ as to maximize:

$$U_j(\mathbf{c}_{j,t}, c_{0j,t}) = \frac{-1}{2\gamma_0} \sum_{i:g_{ij}=0} (c_{ij,t} - a_{ij,t})^2 - \frac{1}{2\gamma_1} \sum_{i:g_{ij}=1} (c_{ij,t} - a_{ij,t})^2 - \frac{1}{2}(c_{0j,t} - a_{0j,t})^2, \tag{3}$$

under the constraint that: $\sum_{i=0}^{n} c_{ij,t} \leq T_{j,t}$, where $T_{j,t}$ is the time constraint for SP j at time t, and $a_{0j,t}$ is the other tasks that SP j needs to do at time t. Therefore, the SP tries to clear their many queues, giving their time constraint and preferences (i.e. $\gamma_0, \gamma_1 > 0$).

Assuming that the time constraint is binding, we can show that:[22]

$$c^*_{ij,t} = a_{ij,t} - \frac{\gamma_{ij}}{\bar{\gamma}_j + 1}(\bar{a}_{j,t} + a_{0j,t} - T_{j,t}), \tag{4}$$

[21] Quantities are all expressed in terms of the opportunity cost of the time required to see a referred patient.

[22] If the time constraint is not binding, then the SP can empty all of their queues. This is highly unrealistic given multiple reports of long and growing queues. For simplicity, we also assume an interior solution (i.e. $c_{ij,t} > 0$), which does not affect the interpretation of the resulting estimating equations. Mathematical derivations can be found in Appendix C.3. See in particular Appendix C.3.1.

where $\bar{\gamma}_j = \sum_{i=1}^{n} \gamma_{ij} = n_j\gamma_1 + (n - n_j)\gamma_0$ and $\bar{a}_{j,t} = \sum_{i=1}^{n} a_{ij,t}$.

The term in parenthesis in Equation (4) is the shortage of medical services: it represents the additional time the SP would need in order to empty all of their queues, and to perform all of their other tasks. At time t, the SP has no control over that shortage. Their only choice is how to allocate the shortage between the n queues of patients. Equation (4) states that the part that is allocated to patients referred by GP i is given by the ratio $\frac{\gamma_{ij}}{\bar{\gamma}_j+1} \in (0,1)$. Thus, if γ_{ij} is relatively *high*, then SP j will see relatively *less* patients referred by i (everything else equal).

While Equation (4) is intuitive, queues are not observed in our data so it is not possible to estimate Equation (4) directly. However, using the queuing process in Equation (2), we can show (see Appendix C.3.2) that:

$$c^*_{ij,t} = \underbrace{b_{ij,t-1}}_{\text{New referrals}} - \underbrace{\gamma_{ij}\dot{\bar{c}}^*_{j,t}}_{\substack{\text{Variation in new} \\ \text{patients seen}}} \underbrace{-\gamma_{ij}[(a_{0j,t} - T_{j,t}) - (a_{0j,t-1} - T_{j,t-1})]}_{\substack{\text{Variation in opportunity cost} \\ \text{of new patient consults}}}, \tag{5}$$

where $\dot{\bar{c}}^*_{j,t} = \bar{c}^*_{j,t} - \bar{c}^*_{j,t-1}$.

Equation (5) allows to express the amount of patient seen from the queue ij at time t as a function of the new referrals at $t-1$, the variation in new patient seen by SP j between t and $t-1$ (an autoregressive term, which is observed) and the variation in the opportunity cost of seeing new patients between t and $t-1$. This specification allows to impose identification assumptions (formalized in Assumption 2) on the primitives of the model.[23]

Assumption 2. *Assume that:*

1. $\mathbb{E}[b_{ij,t-1}] = \alpha^b_0 + \alpha^b_1 n_j$ *for all* $i : g_{ij} = 0$;
2. $\mathbb{E}[b_{ij,t-1}] = \beta^b_0 + \beta^b_1 n_j$ *for all* $i : g_{ij} = 1$;
3. $\mathbb{E}[(a_{0j,t} - T_{j,t}) - (a_{0j,t-1} - T_{j,t-1})] = \delta_0 + \delta_1 n_j$.

Points 1 and 2 state that the expected amount of new referrals are linear functions of the number of links a SP has. Point 3 states that the expected variation in the time available to see referred patients is a linear function of the number of links the SP has. Under Assumption 2, we can rewrite Equation (5) separately for linked and unlinked pairs:

$$c^0_{ij,t} = \alpha_0 + \alpha_1 n_j - \gamma_0 \dot{\bar{c}}_{j,t} + \varepsilon^0_{ij,t}, \tag{6}$$

[23] Contrary to the queues $a_{ij,t}$, which are in part the result of previous choices by SPs, $b_{ij,t}$, $a_{0j,t}$ and $T_{j,t}$ are not functions of these previous choices. It is possible to also allow for linear trends in Assumption 2. In that case, identification would require taking the first difference in equations (6) and (7) below. The absence of sufficient temporal variation in our data however led us to move away from this strategy.

for all $i : g_{ij} = 0$, where $\alpha_0 = \alpha_0^b - \delta_0$, $\alpha_1 = \alpha_1^b - \delta_1$, and

$$c_{ij,t}^1 = \beta_0 + \beta_1 n_j - \gamma_1 \dot{\bar{c}}_{j,t} + \varepsilon_{ij,t}^1, \tag{7}$$

for all $i : g_{ij} = 1$, where $\beta_0 = \beta_0^b - \delta_0$ and $\beta_1 = \beta_1^b - \delta_1$.

Specifications (6) and (7) make it clear that we cannot identify effects coming from more referred patients (increased demand) from effects coming from changes in the resource constraint (supply changes). This is important for the interpretation of the parameters, but does not affect our main parameters of interest, i.e. γ_0 and γ_1.

Equations (6) and (7) help to understand why the event study graphs show changes in the slope after the treatment. Indeed, putting the two equations together, we immediately find:[24]

$$\bar{c}_{j,t} = \psi_j + \frac{\bar{\gamma}_j}{1 + \bar{\gamma}_j} \bar{c}_{j,t-1} + \bar{\varepsilon}_{j,t}, \tag{8}$$

which is an autoregressive process. As discussed above, this is due to the fact that the queues are accumulating at each period. This implies that any treatment on ψ_j will have spillover effects through time: because demand for medical services are not met at each period, the treatment effect accumulates over time. Specifically, the expected impact of a permanent, unanticipated, shock Δ at time t on the medical services rendered by SP j at time $T \geq t$ is:[25]

$$\bar{c}_{j,T}^*(\Delta) - \bar{c}_{j,T}^*(0) = \Delta \sum_{\tau=t}^{T} \left(\frac{\bar{\gamma}_j}{1 + \bar{\gamma}_j} \right)^{\tau - t}.$$

Since queues accumulate through time, the treatment effect is not constant and increases over time. Note however that this equation cannot be directly used to evaluate the treatment effect of the introduction of the CRDS. Indeed, although the argument for the accummulation of the treatment over time is similar, in general, the CRDS does not only impact ϕ_j but affects the entire structure of the referral process. In particular, after the introduction of the CRDS, all SPs of the same specialty who *choose* to use the CRDS essentially share the same queue, which leads to well known free-riding problems.

We now turn to the estimation of equations (6) and (7). If the network was observed, we could estimate both equations by GMM using lagged autoregressive terms as instruments. Since the network is not observed, we use the estimated network from Section 5. This requires checking

[24] Define: $\psi_j(1 + \bar{\gamma}_j) = n(\alpha_0 + \alpha_1 n_j) + n_j(\beta_0 - \alpha_0 + (\beta_1 - \alpha_1) n_j)$, and $\bar{\varepsilon}_{j,t}(1 + \bar{\gamma}_j) = \sum_{i:g_{ij}=0} \varepsilon_{ij,t}^0 + \sum_{i:g_{ij}=1} \varepsilon_{ij,t}^1$.
[25] See Appendix C.3.3.

that Assumption 1 of Section 5 holds for this model.

Proposition 2. *Suppose that Assumption 2 holds and consider equations (6) and (7). Assume that* $(\alpha_0 - \beta_0) + (\alpha_1 - \beta_1)n_j < 0$ *for all* j*. Then, Assumption 1 in Section 5 holds for* $y_{ij} = c^*_{ij}$*.*

If the arrival of new referrals $b_{ij,t}$ and changes in the effective time constraint are, in expectation, larger when i and j are linked, then the ordering is also preserved in the amount of new patients seen (also in expectation). Proposition 2 therefore provides sufficient conditions so that the structural model presented in this section is coherent with the assumption under which our estimator of the network is consistent.

We can therefore estimate equations (6) and (7) conditional on our estimated network structure. Results are presented in Table 4.[26]

Not linked			**Linked**		
Variables	Estimate	Std.Error	Variables	Estimate	Std.Error
γ_0	0.0022	0.0000	γ_1	0.1496	0.0502
α_0	0.8445	0.0017	β_0	53.2443	1.6150
α_1	0.0326	0.0007	β_1	-1.6717	0.2467

Note: Left panel gives estimated values and standard errors for the model in Equation (6). Right panel gives estimated values and standard errors for the model in Equation (7). Both equations were estimated independently using two stage least squares, with $\dot{\bar{c}}_{t-2}$ as instrument for $\dot{\bar{c}}_{t-1}$ to address the simultaneity with the explained variable.

Table 4: Results from the structural model

Unsurprizingly, the expected amount of new referrals is much larger for linked pairs. For pairs of physicians that are not linked, the expected amount of new referrals is slightly increasing in the number of links that the SP has. Recall that $\alpha_1 = \alpha_1^b - \delta_1$, where δ_1 is the variation of the expected time available to the SP, as a function of the number of link they have. We therefore interpret $\alpha_1 > 0$ as evidence that $\delta_1 < 0$: SPs with large number of links see a decrease (or a lesser growth) in the time available to see new patients. For linked pairs, the expected amount of new referrals is decreasing in the number of links the SP has, which suggests a congestion effect.[27]

We now turn to our main parameters of interest, i.e. γ_0 and γ_1. We first note that the estimated value for γ_1 is much larger than that for γ_0. This implies that the utility cost of any given queue of patients is lower if patients were referred by a linked GP. Note however that these numbers have to be analyzed in context. Indeed, there are *much more* unlinked pairs of

[26] Using the estimated parameters in Table 4, the condition in Proposition 2 holds for all SPs in our sample.
[27] For example, linked GP may receive information from previously referred patients about the SP's workload.

physicians (and thus much more queues from unlinked GPs). We also expect queues of patients referred by a linked GP to be much longer (see below).

Thus, a measure that is perhaps more intuitive is the ratio of the medical services over the length of the queue:

$$\frac{c^*_{ij,t}}{a_{ij,t}}.$$

We can show that this quantity is larger for the linked pair ij than the unlinked pair kj iff[28]

$$\frac{\gamma_1}{\gamma_0} \leq \frac{a_{ij,t}}{a_{kj,t}}.$$

When $\gamma_1 = \gamma_0$, SP j provides relatively more medical services for queue ij iff it is longer. When $\gamma_1 > \gamma_0$, as implied by Table 4, SP j provides relatively more medical services to queue ij only if the queue is much longer than queue kj.

But, because i and j are linked, queue ij is expected to be longer. Indeed, under Assumption 2, and after some algebra, we find that SP j provides relatively more medical services to queue ij (with a linked GP) relative to queue kj (with a unlinked GP) iff:[29]

$$\gamma_1[\alpha_0 + \alpha_1 n_j + \eta_{kj,t}] \leq \gamma_0[\beta_0 + \beta_1 n_j + \eta_{ij,t}].$$

At the estimated values, this condition rejected by the data for all SP.[30]

This implies that the fact that the utility cost of queues for SPs is smaller when patients are referred by a linked GP, implies that SP provide relatively less medical services to patients referred by linked GP.

7 Conclusion

In this paper, we leverage detailed physicians' billing data to find that the introduction of regional centralized referral systems (CRDS) significantly reduced the average volume of new referrals from GPs seen by SPs, especially among high-referral physicians. SPs appear to reallocate their time toward their existing patient roster, and toward more complex medical procedures, particularly in hospital settings. These results are found despite the relatively low adoption rate of the CRDS. While our data does not allow us to directly observe queues and waiting times,

[28] See Appendix C.3.4

[29] See Appendix C.3.4

[30] Replacing values with the estimated ones in Table 4, and letting $\eta_{ij,t} = 0$, we find $\min_j \hat{\gamma}_1[\hat{\alpha}_0 + \hat{\alpha}_1 n_j]/\hat{\gamma}_0[\hat{\beta}_0 + \hat{\beta}_1 n_j] = 1.08$. Note that the denominator is positive for all SP in the sample.

these results are at odds with an objective of improving access and reducing the length of the period between the initial referral is filed and the moment a patient sees a physician specialist, a goal associated with potentially improved health and labour market outcomes for patients (e.g. Godøy et al. (2024), Siciliani (2026)). One area where CRDS might have generated welfare improvements is on the prioritization of the patients effectively seen by SPs post referrals (Gravelle and Siciliani, 2008) – This question is left for future research as our data does not enable us to investigate it.

Using a structural model estimated on pre-CRDS data, we then find that while specialists dislike long queues of patients waiting, they tend to defer consultations for patients referred by general practitioners from whom they receive an unusually large number of referrals. We interpret this finding as evidence that SPs value equity of patient access.

Our results on the reduced number of consultations contrasts with the existing medical literature. We cannot know if this difference is due to differences in methodology used, or to the specificity of the CRDS. That said, the low adoption rate, as well as informal discussion with SPs leads us to believe that while the CRDS aggregates information well, it is also not very flexible. Accepting patients from the CRDS may also represent a *riskier* decision for SPs, who may be more unsure about this patient's likelihood to show up to their appointment, the quality of the GPs' assessment, etc. Thus, although this centraliztion of waiting lists probably improved the equality of access for patients (which is also valued by SPs), the welfare impacts are ambiguous: the reallocation of SPs service allocation may be welfare improving, or it may be the result of inefficiencies created by the CRDS. In particular, our results highlight the importance of accounting for administrative burden and rigidities of centralizing referral systems. Indeed, in the context of the CRDS, these administrative costs likely contribute to the low adoption rate. The loss of private information shared through referral networks might also affect physician behaviour, including their adoption of centralized mechanisms, and their allocation of services post-adoption.

Studying further the impact of the CRDS would require detailed data of the system itself. In particular, we suspect important free-riding issues: with the CRDS, SP essentially share the same queue of patients, which could result in under-provision of availabilities on the part of the SP. These considerations are left for future research.

References

Batabyal, A. A. and H. Beladi (2008): "Bribery and favoritism in queuing models of rationed resource allocation," *Journal of Theoretical Politics*, 20, 329–338.

Bonhomme, S. (2025): "Back to feedback: Dynamics and heterogeneity in panel data," *arXiv preprint arXiv:2512.17576*.

Brosig-Koch, J., M. Griebenow, M. Kifmann, and F. Then (2022): "Rewards for information provision in patient referrals: A theoretical model and an experimental test," *Journal of Health Economics*, 86, 102677.

Callaway, B. and T. Li (2019): "Quantile treatment effects in difference in differences models with panel data," *Quantitative Economics*, 10, 1579–1618.

Clair, M. (2000): "Commission d'étude sur les services de santé et les services sociaux," *Les solutions émergentes*.

Cornwall, G. and S. Wentland (2025): "It's About Time (Series): A Simple Correction For Difference-in-Differences Estimators," .

Damani, Z., B. Conner-Spady, T. Nash, H. Tom Stelfox, T. W. Noseworthy, and D. A. Marshall (2017): "What is the influence of single-entry models on access to elective surgical procedures? A systematic review," *BMJ open*, 7, e012225.

de Carvalho Nunes, L. F., F. J. M. D. Costa, and F. A. M. Sanches (2016): "PRACTICE LOCATION OF PHYSICIANS: A DISCRETE CHOICE MODEL APPROACH," Tech. rep., ANPEC-Associa.

De Santana, V. F., A. P. Appel, L. G. Moyano, M. Ito, and C. S. Pinhanez (2018): "Revealing physicians referrals from health insurance claims data," *Big data research*, 13, 3–10.

Di Renna, T., E. Burke, A. Bhatia, H. Clarke, D. Flamer, J. Flannery, A. Furlan, D. Kumbhare, J. Khan, K. Ladha, et al. (2024): "Improving access to chronic pain care with central referral and triage: the 6-year findings from a single-entry model," *Canadian Journal of Pain*, 8, 2297561.

for Health Information", C. I. (2022): "An overview of physician payments and cost per service," Accessed: 2026-09-01.

Godøy, A., V. F. Haaland, I. Huitfeld, and M. Votruba (2024): "Hospital Queues, Patient Health, and Labor Supply," *American Economic Journal: Economic Policy*, 16, 150–181.

Gravelle, H. and L. Siciliani (2008): "Is waiting-time prioritization welfare-improving?" *Health Economics*, 17, 167–184.

Greenwood-Lee, J., L. Jewett, L. Woodhouse, and D. A. Marshall (2018): "A categorisation of problems and solutions to improve patient referrals from primary to specialty care," *BMC health services research*, 18, 986.

Hackl, F., M. Hummer, and G. J. Pruckner (2015): "Old boys' network in general practitioners' referral behavior?" *Journal of health economics*, 43, 56–73.

Jackson, M. O., B. W. Rogers, and Y. Zenou (2017): "The economic consequences of social-network structure," *Journal of Economic Literature*, 55, 49–95.

Kennedy, L., B. O. Asante, and F. Clement (2025): "Patient and provider perspective with the use of a central intake system (CIS) for surgical waitlist management: a systematic review," *BMJ open*, 15, e091530.

Milakovic, M., A. M. Corrado, M. Tadrous, M. E. Nguyen, S. Vuong, and N. M. Ivers (2021): "Effects of a single-entry intake system on access to outpatient visits to specialist physicians and allied health professionals: a systematic review," *CMAJ open*, 9, E413–E423.

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

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

Roth, J., P. H. Sant'Anna, A. Bilinski, and J. Poe (2023): "What's trending in difference-in-differences? A synthesis of the recent econometrics literature," *Journal of Econometrics*, 235, 2218–2244.

Sarsons, H. (2017): "Interpreting signals in the labor market: evidence from medical referrals," *Job Market Paper*, 141–45.

Siciliani, L. (2026): "Waiting times for health services, health, and labour market outcomes," *European Journal of Public Health*, 36, ii3–ii7.

Veinot, T. C., E. A. Bosk, K. Unnikrishnan, and T. J. Iwashyna (2012): "Revenue, relationships and routines: The social organization of acute myocardial infarction patient transfers in the United States," *Social science & medicine*, 75, 1800–1810.

Vo, A. T., Y. Yi, M. Mathews, J. Valcour, M. Alexander, and M. Billard (2022): "A single-entry model and wait time for hip and knee replacement in eastern health region of Newfoundland and Labrador 2011–2019," *BMC Health Services Research*, 22, 82.

Zeltzer, D. (2020): "Gender homophily in referral networks: Consequences for the medicare physician earnings gap," *American Economic Journal: Applied Economics*, 12, 169–97.

A Data – Additional information

Our estimating sample contains physician billing data from ten administrative regions in the province of Québec (generally the most populous). For the heterogeneity analysis and the structural estimation, we pool the data into three broader regions (based on geographic proximity and, for groups 1, 2 and 3, on existing interactions).

Region 1: Montréal, Laval, Lanaudière, Laurentides, Montérégie

Region 2: Québec (Capitale-Nationale), Mauricie, Centre-du-Québec, Chaudière-Appalaches

Region 3: Estrie

Our estimating sample includes data for specialties for which the CRDS was either implemented in 2016 or *not yet* implemented in 2019.

CRDS implementation phases

- The first phase was in the fall of 2016 and affected referrals to the specialties: Cardiology, Gastroenterology, Nephrology, Neurology, otorhinolaryngology (ORL), Ophthalmology, Orthopedics, Pediatrics, and Urology;
- The second phase was in the fall of 2018 and affected referrals to the specialties: Internal medicine, Physiatry, Rheumatology, Neurosurgery, General Surgery, Hemato-Oncology, and Vascular Surgery;
- The third phase was planned in the fall of 2019 and affected referrals to the specialties: Pneumology, Obstetrics and Gynecology, Allergy and Clinical Immunology, Plastic Surgery, Endocrinology, Psychiatry, Geriatrics, Dermatology, and Medical Microbiology-Infectiology.

Our estimating sample is constituted of the specialties below. Following informal discussions with physicians, we excluded some treated and untreated specialties. Specifically: Diagnostic Radiology (these are more diagnostic services, and not proper consultations with the patients); Psychiatry, and Thoracic and Vascular surgery (highly volatile and specialized, not often directly referred by the GP and hardly comparable with the other specialties.)

Estimation of the adoption

Our data, which ends in 2019, allows us to compute the number of consultations at any given period (say, daily, weakly...). While each SP billing record for a new consultation identifies

Code	Specialty	Treatment status
6	Cardiology	Treated
9	Plastic Surgery	Not treated
11	Dermatology	Not treated
12	Gastroenterology	Treated
17	Pneumology	Not treated
25	ORL	Treated
27	Pediatrics	Treated
31	Urology	Treated

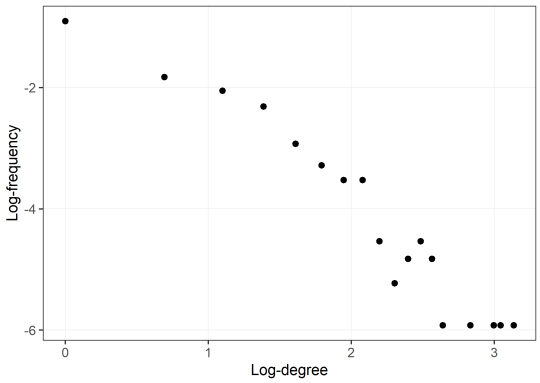
Table A.1: List and codes of Specialties in our sample

(through an anonymized number) the referring GP, we do not know if it was a direct referral, or a referral that went through the CRDS. However, starting 2021, the CRDS is publishing the number of consultations that went through the CRDS.[31] Assuming that the the number of consultations was roughly stable for the period 2016–2019 and 2021–2026 for our treated specialties, we can compute the fraction of consultations for which the referral went through the CRDS: 46% (Urology), 18% (Pediatrics), 35% (ORL), 12% (Gastroenteroloy), and 16% (Cardiology). These numbers are compatible with numbers appearing in the press, in which both the *Fédération des médecins spécialistes du Québec* (Association of SPs) and the minister of health state that 20% of referrals go through the CRDS.[32]

[31] See https://www.quebec.ca/sante/systeme-et-services-de-sante/organisation-des-services/donnees-systeme-sante-quebecois-services/performance-reseau-sante-services-sociaux. (In French)

[32] See https://ici.radio-canada.ca/nouvelle/2239292/medecins-SPes-fmsq-nouveau-moyen-pression (In French) and https://www.ledevoir.com/actualites/sante/964086/medecins-SPes-entament-nouveau-moyen-pression (In French).

A.1 Degree distribution of the estimated referral network



Note: This figure displays the log-log degree distribution, conditional on a positive degree. The approximate linearity is evidence of a fat-tail distribution.

Figure A.1: Fat-tail degree distribution

B Reduced-form estimates: complementary analysis

B.1 Difference-in-difference estimates

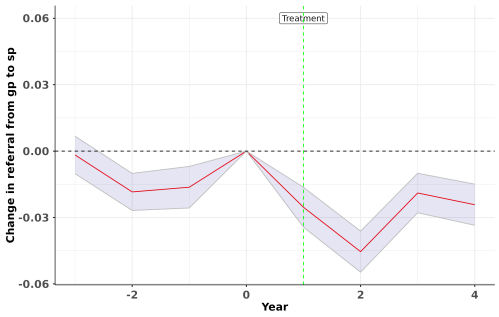
All specifications point to a negative treatment effect on the volume of new consultations, both at the SP level and at the GP–SP pair level. In the baseline two-way fixed-effects specification, the treatment effect is negative and statistically significant for both outcomes. When allowing for a differential linear trend in Specification B, the estimated post-treatment effect remains negative but is no longer statistically significant at the SP level, while it remains negative and significant at the pair level. The differential linear trend itself is negative and statistically significant in both samples, which is consistent with the patterns observed in the event-study graphs. Finally, Specification C includes an autoregressive term, which is positive and highly significant for both SPs and pairs. This indicates persistence in referral patterns over time and is consistent with the dynamic structure of our structural model, in particular Equation (8).

SPs			
	Specification A	Specification B	Specification C
Treatment effect	−66.399***	−19.346	−38.666***
	(23.341)	(19.294)	(14.234)
Treatment on linear trend	–	−10.456**	–
	–	(5.031)	–
Autoregressive term	–	–	0.573***
	–	–	(0.024)
Observations	16,760	16,760	14,895
Pairs			
	Specification A	Specification B	Specification C
Treatment effect	−0.041***	−0.017***	−0.033***
	(0.004)	(0.004)	(0.002)
Treatment on linear trend	–	−0.005***	–
	–	(0.001)	–
Autoregressive term	–	–	0.394***
	–	–	(0.012)
Observations	15,245,480	15,245,480	13,551,535

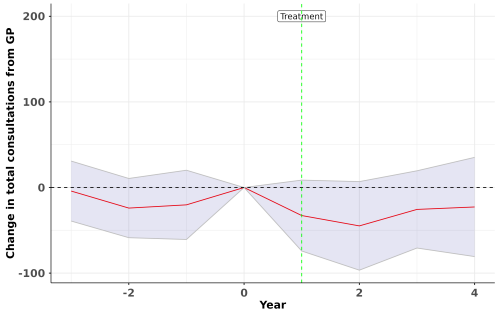
Note: Specification A is a simple two-way fixed-effects difference-in-differences model. Specification B allows for a differential linear trend, while Specification C includes an autoregressive term. All specifications include time and SP/pair fixed effects. Standard errors are clustered at the SP/pair level and reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B.1: Difference-in-differences estimates.

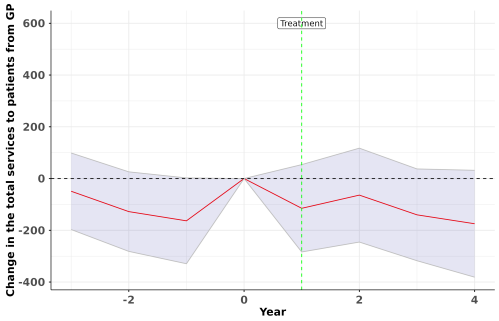
B.2 Event-study graphs on first-differences



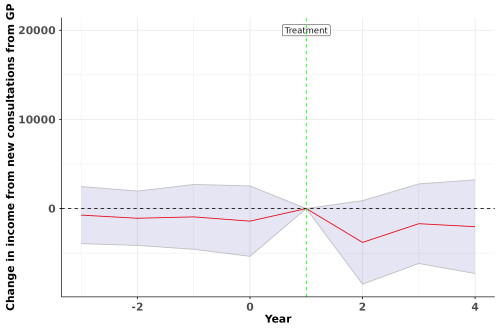
(a) Patient referrals between pair of GP-SP



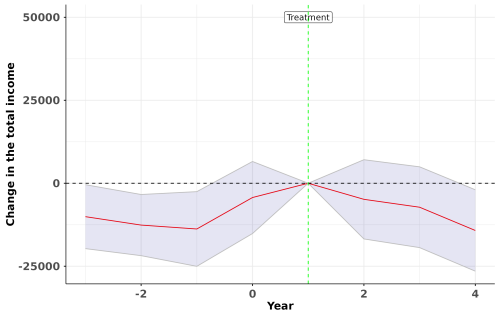
(b) Volume of new referral from GP to SP



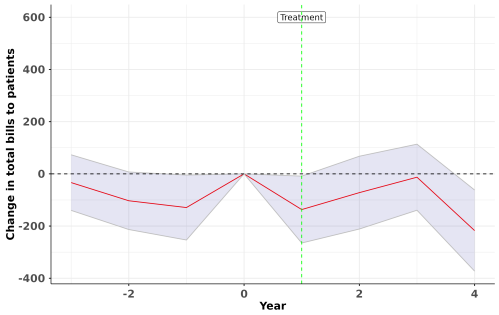
(c) Volume of total services provided by SP



(d) Income from new referral from GP to SP



(e) Income from total services provided by SP



(f) Volume of total invoice billed by SP

Figure B.1: Impact of centralization on outcomes (in variation)

B.3 Sensitivity analysis and robustness checks

Our estimates primarily indicate a significant decline in the average number of patient referrals shared between GP–SP pairs following centralization. To validate our identification assumptions and corroborate these results, we perform a series of robustness checks.

We begin with placebo tests using only the pre-reform period. Specifically, we restrict the sample to observations before the implementation of the CRDS reform and assign fictitious treatment dates in 2013, 2014, and 2015. For each placebo specification, we construct an indicator equal to one for treated specialties after the corresponding fake treatment date. The results, reported in Table B.2, show no statistically significant placebo effect: the estimated coefficients are small in magnitude and statistically indistinguishable from zero across all three fake treatment dates. This absence of significant pre-reform effects supports the identifying assumption that treated and control specialties were not already following divergent trends before centralization. It also suggests that the estimated post-2016 effects are unlikely to be driven by anticipation effects or by pre-existing differential dynamics, but rather by the actual implementation of the CRDS reform.

Number of patients referred by general practitioners (GPs) to SPs			
	Fake Treatment in 2013	Fake Treatment in 2014	Fake Treatment in 2015
Treatment	−0.001	−0.008	0.007
	(0.009)	(0.011)	(0.014)
Num. Obs.	8,469,710	8,469,710	8,469,710
R2	0.785	0.785	0.785
R2 Adj.	0.731	0.731	0.731
AIC	32,924,093.7	32,924,085.3	32,924,087.6
BIC	56,558,053.8	56,558,045.4	56,558,047.7
RMSE	1.38	1.38	1.38
Std. Errors		Clustered by GP–SP pair	
FE: Pair of GP and SP	X	X	X
FE: Year	X	X	X

Note: Each column reports a placebo difference-in-differences estimate using a fake treatment date before the actual implementation of the reform. The dependent variable is the number of patients referred by general practitioners (GPs) to specialists (SPs). All specifications include GP–SP pair fixed effects and year fixed effects. Standard errors are clustered at the GP–SP pair level and reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B.2: Placebo test estimates.

Number of new consultations from GP			
	Fake Treatment in 2013	Fake Treatment in 2014	Fake Treatment in 2015
Treatment	−2.922	−16.843	−5.434
	(9.772)	(12.471)	(15.210)
Num. Obs.	9,310	9,310	9,310
R2	0.932	0.932	0.932
R2 Adj.	0.914	0.914	0.914
R2 Within	0.000	0.000	0.000
R2 Within Adj.	-0.000	0.000	-0.000
AIC	125,612.4	125,609.7	125,612.2
BIC	138,940.6	138,938.0	138,940.4
RMSE	168.45	168.42	168.45
Std. Errors		Clustered by SP	
FE: SP ID	X	X	X
FE: Year	X	X	X

Note: Each column reports a placebo difference-in-differences estimate using a fake treatment date before the actual implementation of the reform. The dependent variable is the number of new consultations from general practitioners (GPs) to specialists (SPs). All specifications include SP fixed effects and year fixed effects. Standard errors are clustered at the SP level and reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B.3: Placebo test estimates at the specialist level.

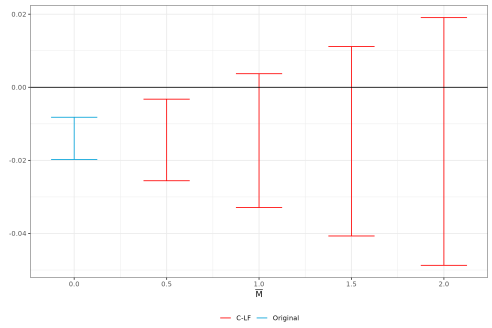
We also conduct sensitivity analyses for the treatment effects in 2016 and 2017, as well as for the average treatment effect of centralization. Two complementary approaches are applied: (Rambachan and Roth, 2023)

(1) Bounds on Relative Magnitudes (BRM): This method assumes that deviations from parallel trends in the post-treatment period cannot be disproportionately larger than those observed before treatment. It introduces a constant $\bar{M}$ such that post-treatment violations are at most $\bar{M}$ times the maximum pre-treatment deviation. For instance, setting $\bar{M} = 1$ ensures that post-treatment violations are no worse than the largest observed pre-trend deviation.

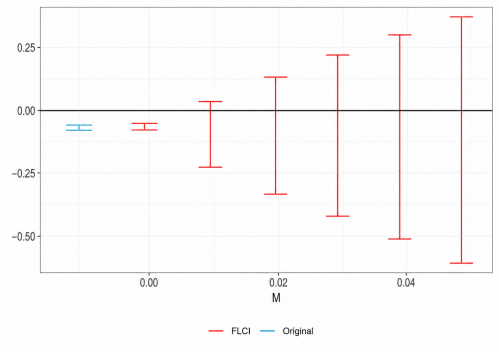
(2) Smoothness Restrictions (SR): This approach constrains post-treatment trend deviations by limiting how sharply they can diverge from the linear extrapolation of the pre-trend. It imposes a bound on the change in slope across consecutive periods, ensuring smooth transition dynamics.

The results in Figure B.2 reveal that the "breakdown value" for a significant effect ranges from $\bar{M} = 0.5$ to $\bar{M} = 1.5$, indicating that the significant findings are robust to violations of parallel trends up to 1.5 times the maximum pre-treatment violation. However, regarding the sensitivity analysis based on the smoothness restriction for the average treatment effect, we see that the breakdown value for a significant effect is $M \approx 0.01$, meaning that we can reject a null

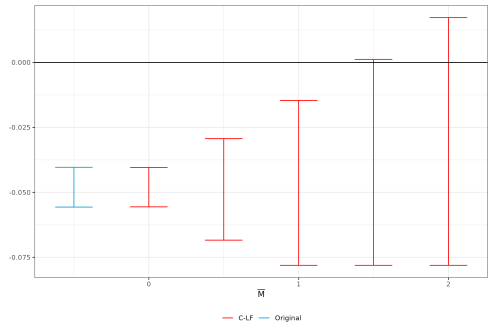
effect unless we are willing to allow the linear extrapolation across consecutive periods to be off by more than 0.01.



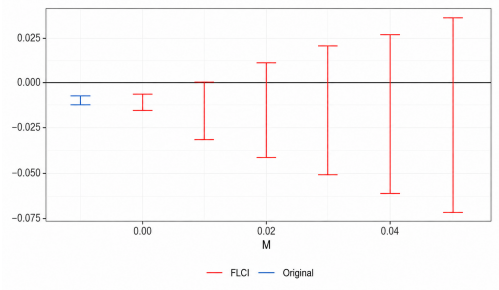
(a) Honest DID for average treatment effect on pair (BRM)



(b) Honest DID for average treatment effect on pair (SR)



(c) Honest DID for treatment effect in 2016 on pair (BRM)



(d) Honest DID for treatment effect in 2016 on pair (SR)

Figure B.2: Sensitivity analyses

B.4 Heterogeneity analysis

To better understand the mechanisms underlying the effects of the CRDS reform, we conduct a series of heterogeneity analyses.

Distributional Effects

We begin by examining distributional effects using Quantile Treatment Effects (QTE). As shown in Figure B.3 and Figure B.4, the effects of centralization are heterogeneous across the distribution. For new consultations from GPs, the estimated QTEs are close to zero over most of the distribution but become more negative in the upper tail. This suggests that the decline in new consultations is concentrated among SPs who were initially receiving relatively high volumes of GP referrals. A similar pattern is observed for the variation in new consultations, where the negative effects are more pronounced at higher quantiles.

For income from new consultations, the effects are less uniform. The estimates suggest negative effects at the lower and upper parts of the distribution, while some mid-distribution quantiles display positive effects. However, the confidence intervals are wider, especially at the tails of the distribution, so these results should be interpreted with caution. For total services and total bills, the QTEs also reveal heterogeneous responses: some mid-distribution SPs appear to experience modest increases, while the upper tail shows weaker or negative effects, particularly in the variation outcomes. Overall, the distributional results are consistent with a redistributive mechanism: centralization reduced activity among high-referral SPs, while some lower- or middle-volume SPs may have absorbed part of the redirected demand.

Effects by Type of Medical Establishment

We also explore heterogeneity in treatment effects across types of medical settings, comparing hospital-based SPs to those practicing in other centers, such as private practices or clinics. The results in Figure B.5 show that the decline in the average number of patients shared between GP–SP pairs is visible in both hospital centers and other centers. In hospital centers, the post-reform effect becomes increasingly negative over time, both in levels and in growth rates. A similar post-reform decline is observed in other centers, although the confidence intervals suggest that some estimates are less precisely measured.

The patterns for new consultations provided by SPs are reported in Figure B.6. In hospital centers, the estimated effects on new consultations decline after the reform, but the confidence intervals are relatively wide. In other centers, the post-reform effects are also negative, especially in levels, suggesting that the reduction in GP-originating consultations was not limited to hospital-based interactions. These results indicate that the CRDS affected referral relationships

across different types of medical settings, rather than exclusively through hospitals.

Turning to total services, Figure B.7 suggests a more nuanced pattern. Among hospital-based SPs, total services appear to increase modestly after the reform, although the estimates are imprecise. In contrast, SPs practicing in other centers display more negative post-reform effects, especially in levels. This difference suggests that while the CRDS reduced new GP-originating consultations, hospital-based SPs may have partially compensated through other types of services, whereas SPs outside hospitals experienced a broader decline in service volumes.

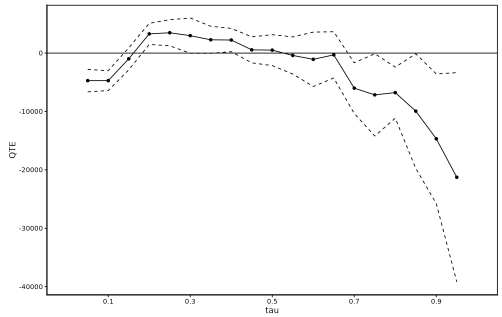
Effects by region

Finally, we examine whether the effects of the CRDS differ across regions of Quebec. To account for potential interregional referrals and to ensure sufficient sample size within each subgroup, we pool the 10 health regions into three larger regional groups, as described in Appendix A.

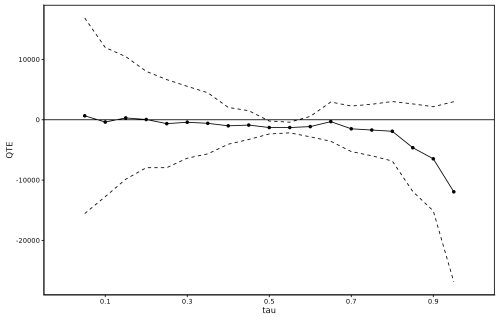
The results in Figure B.8 show that the decline in the average number of patients shared between GP–SP pairs is not uniform across regions. Region 1 displays a clear post-reform decline in both levels and growth rates. Region 2 shows smaller and more stable effects, with estimates remaining closer to zero. Region 3 exhibits a sharper negative response shortly after the reform, although the pattern is less persistent over time. These differences suggest that the impact of centralization on referral relationships varied across local health-care markets.

For new consultations provided by SPs, shown in Figure B.9, the regional patterns are more moderate. Region 1 shows a gradual post-reform decline, while Regions 2 and 3 display weaker and less precisely estimated effects. For total services, reported in Figure B.10, the confidence intervals are generally wide across all three regional groups, and the estimates do not point to a clear statistically robust pattern. This suggests that regional heterogeneity is more pronounced for referral relationships than for overall service volumes.

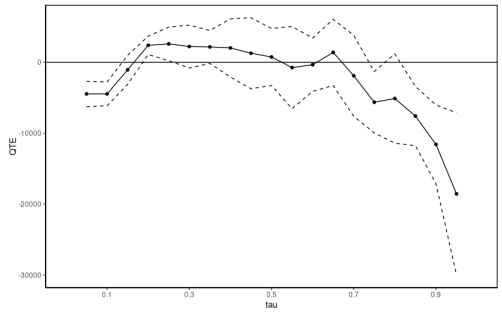
Taken together, the heterogeneity analyses indicate that the CRDS mainly affected established GP–SP referral relationships, with stronger effects among high-referral SPs and in some regional and institutional settings. The evidence is consistent with a redistribution of referrals away from concentrated pre-existing relationships, although the effects on total services are more heterogeneous and less precisely estimated. While these analyses provide useful descriptive evidence on where the effects are concentrated, they offer limited insight into the underlying decision-making behavior of physicians. To address this limitation, we develop a structural model in the next section that allows us to investigate how GPs and SPs adjust their referral and consultation choices in response to the introduction of the CRDS.



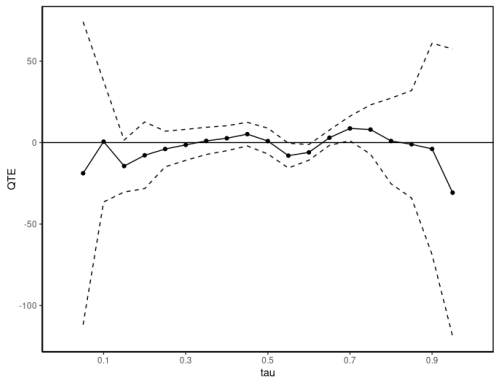
(a) QTE on new consultations (Level)



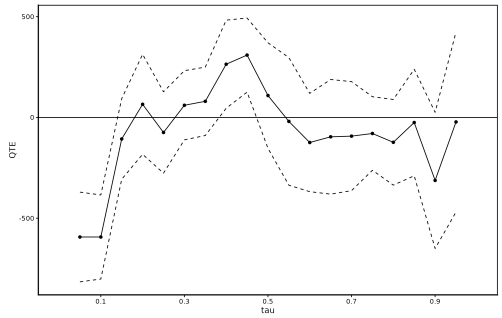
(b) QTE on new consultations (Variation)



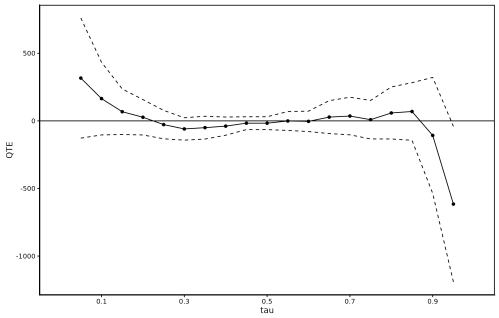
(c) QTE on income from new consultations (Level)



(d) QTE on income from new consultations (Variation)

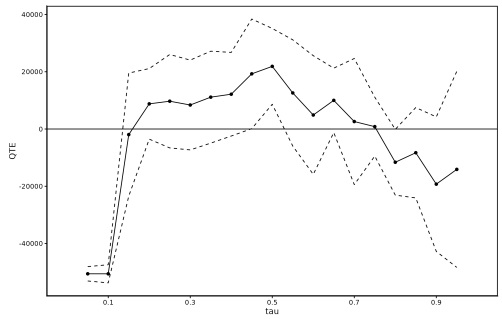


(e) QTE on total services (Level)

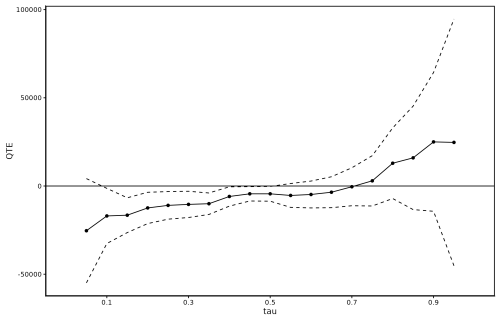


(f) QTE on total services (Variation)

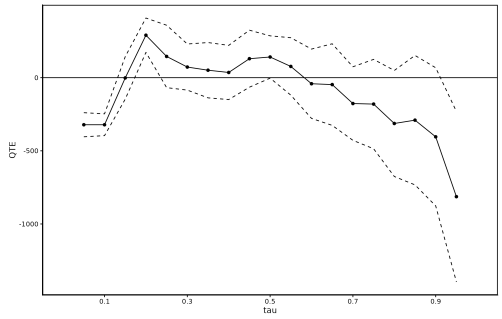
Figure B.3: Heterogenous effect: Quantile Treatment Effects



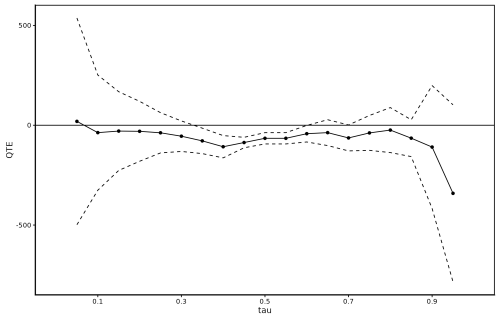
(a) QTE on total income (Level)



(b) QTE on total income (Variation)

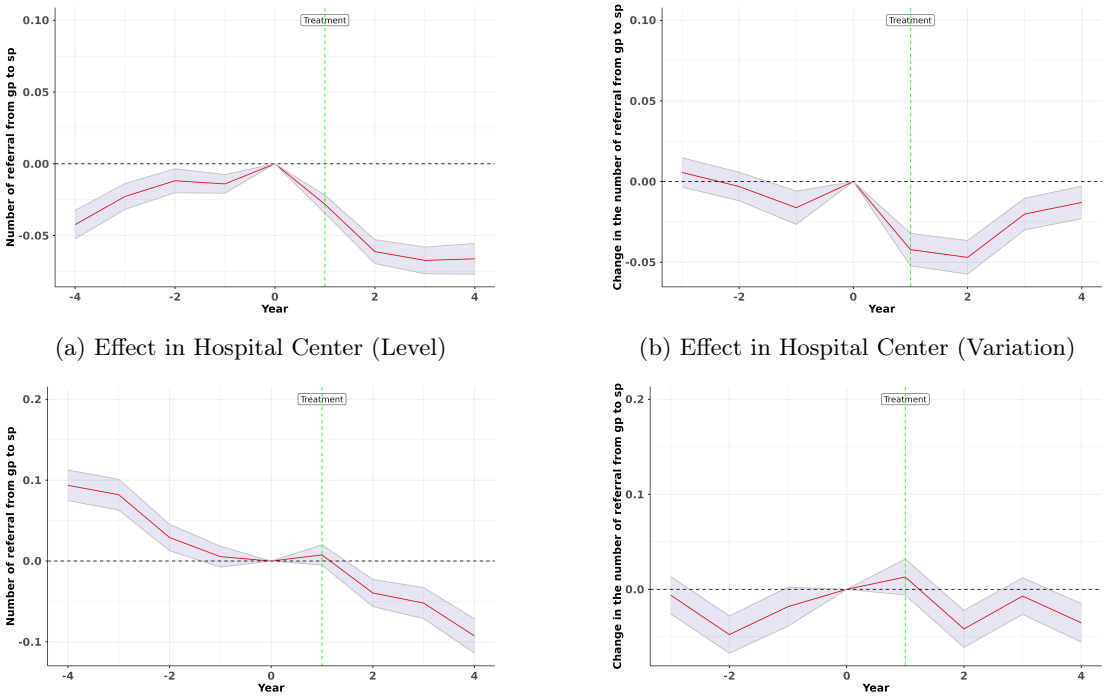


(c) Quantile Treatment Effects on total bills (Level)



(d) Quantile Treatment Effects on total bills (Variation)

Figure B.4: Heterogenous effect: Quantile Treatment Effects



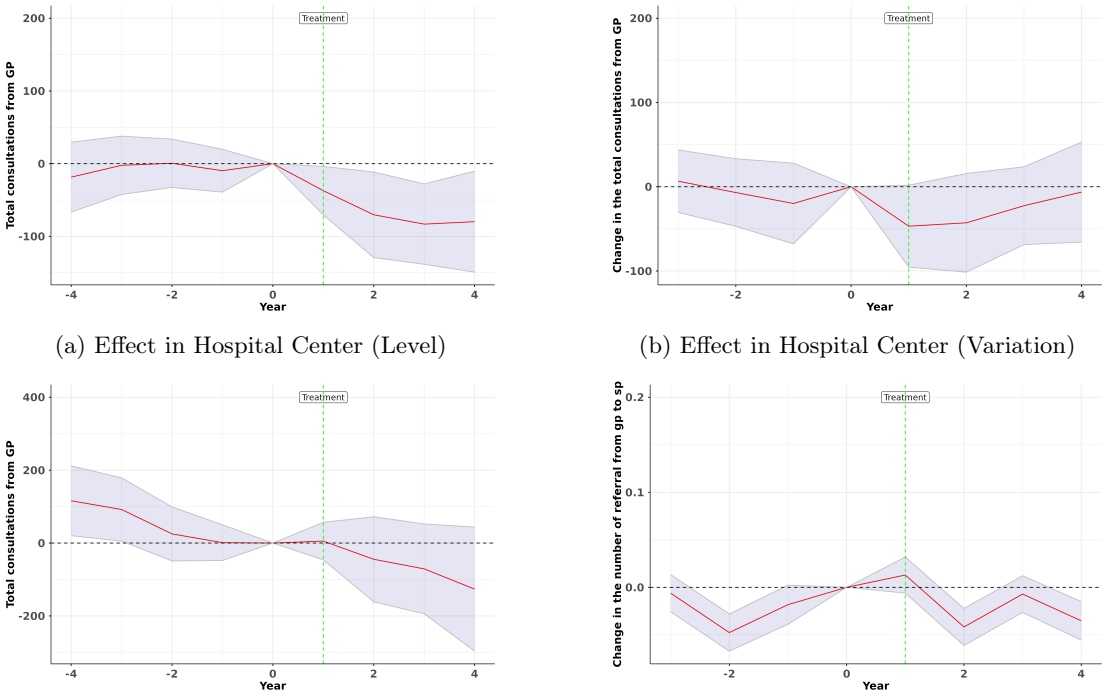
(a) Effect in Hospital Center (Level)

(b) Effect in Hospital Center (Variation)

(c) Effect in other centers (Level)

(d) Effect in other centers (Variation)

Figure B.5: Heterogenous effect on the number of patients between pairs of SPs (SPs) and general practitioners (GPs) per type of establishment



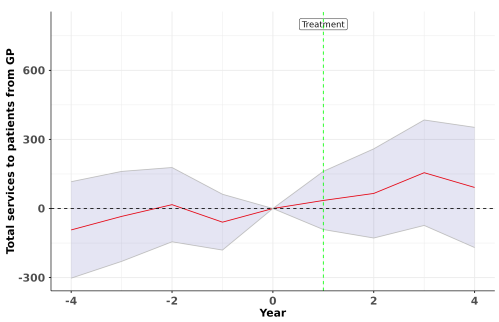
(a) Effect in Hospital Center (Level)

(b) Effect in Hospital Center (Variation)

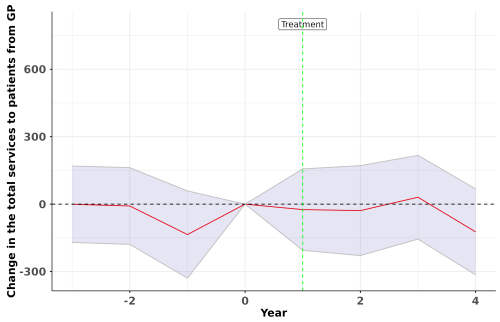
(c) Effect in other centers (Level)

(d) Effect in other centers (Variation)

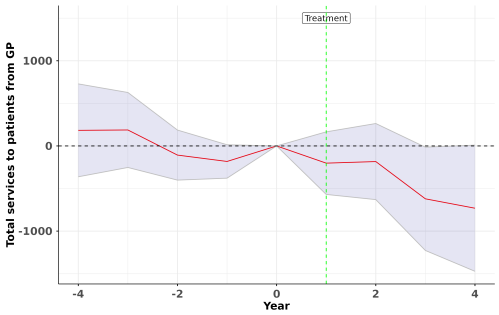
Figure B.6: Heterogenous effect on the number of new consultations provide by SP per type of establishment



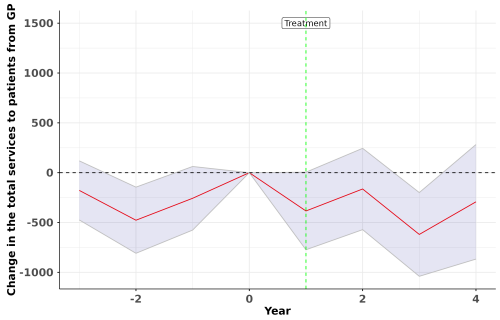
(a) Effect in Hospital Center (Level)



(b) Effect in Hospital Center (Variation)

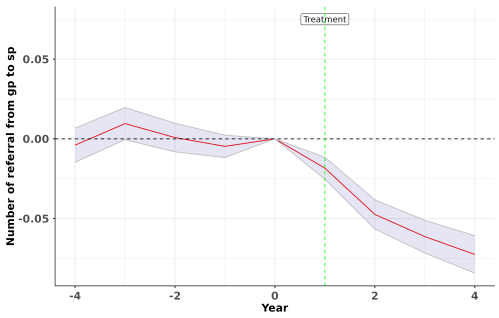


(c) Effect in other centers (Level)

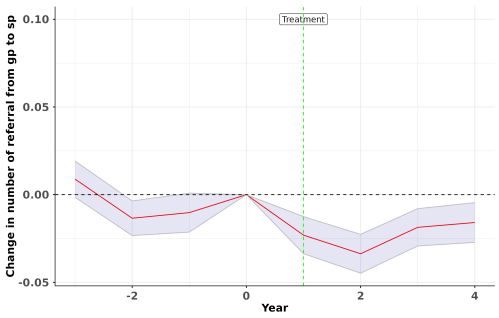


(d) Effect in other centers (Variation)

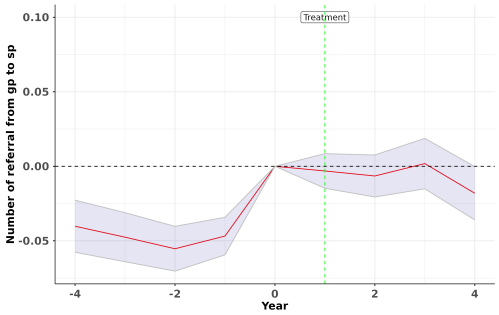
Figure B.7: Heterogenous effect on the total services given by SP per type of establishment



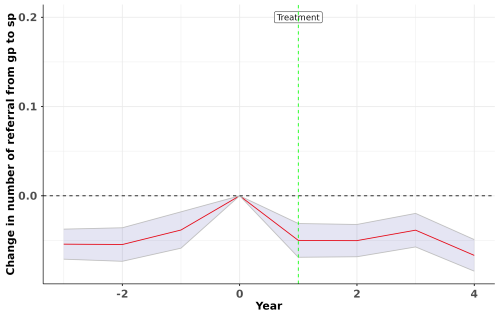
(a) Effect in region 1 - Montreal and neighbouring regions (Level)



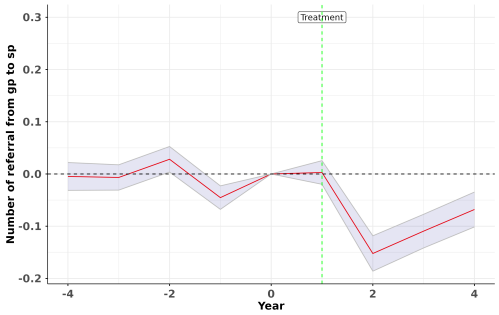
(b) Effect in region 1 - Montreal and neighbouring regions (Variation)



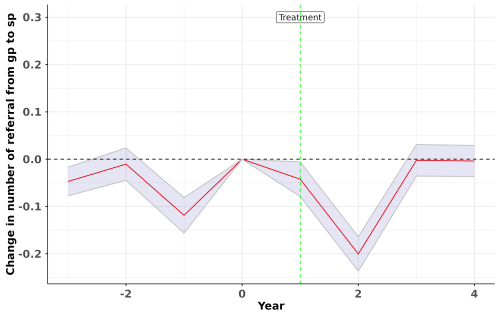
(c) Effect in region 2 - Quebec and neighbouring regions (Level)



(d) Effect in region 2 - Quebec and neighbouring regions (Variation)

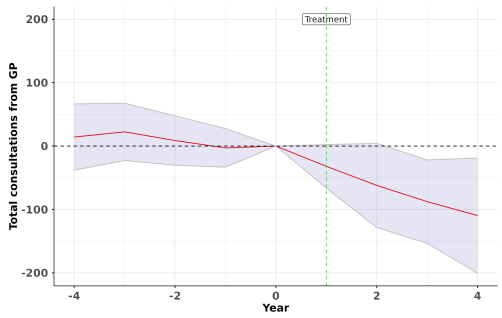


(e) Effect in region 3 - Estrie (Level)

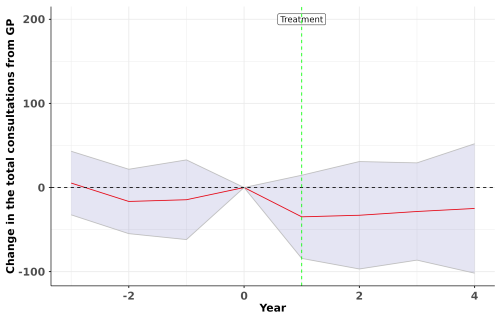


(f) Effect in region 3 - Estrie (Variation)

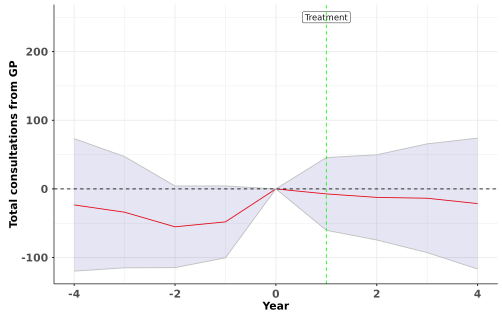
Figure B.8: Heterogenous effect on the number of patients between pairs of SPs (SPs) and general practitioners (GPs) per group of regions



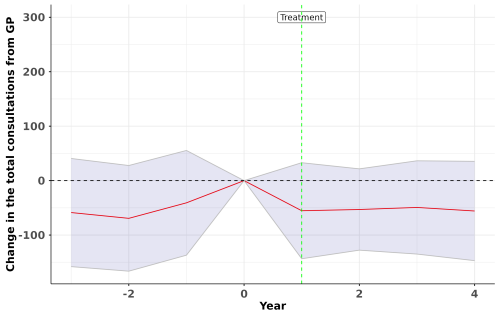
(a) Effect in region1 - Montreal and neighbouring regions (Level)



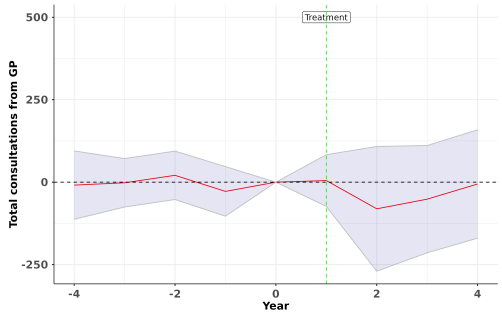
(b) Effect in region 1 - Montreal and neighbouring regions (Variation)



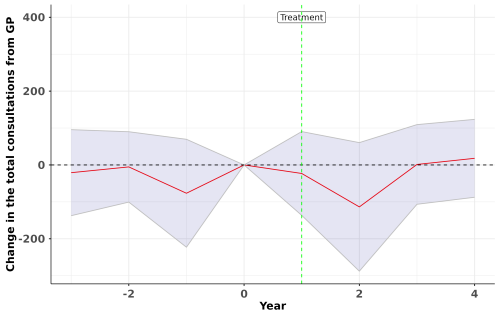
(c) Effect in region 2 - Quebec and neighbouring regions (Level)



(d) Effect in region 2 - Quebec and neighbouring regions (Variation)

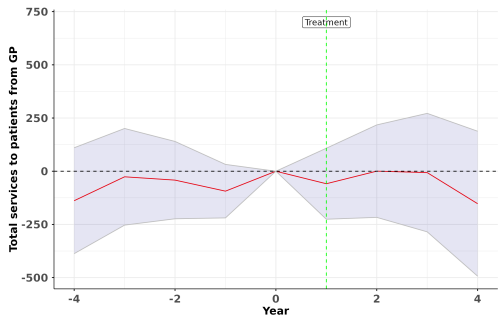


(e) Effect in region 3 - Estrie (Level)

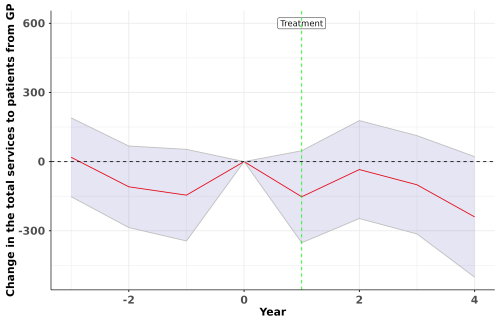


(f) Effect in region 3 -Estrie (Variation)

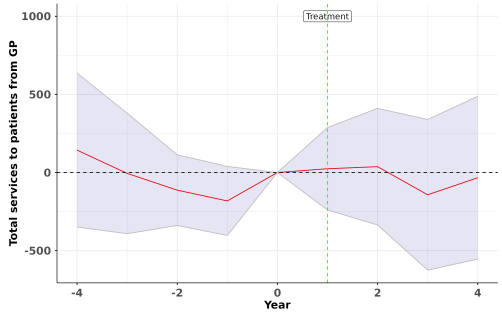
Figure B.9: Heterogenous effect on the number of new consultations provide by SP per group of regions



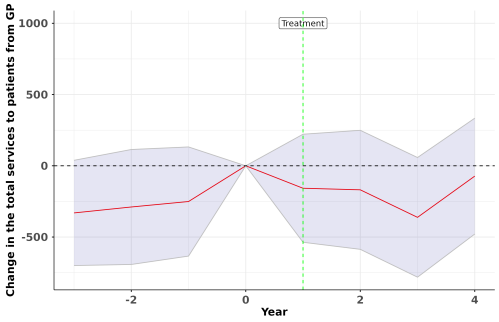
(a) Effect in region1 - Montreal and neighbouring regions (Level)



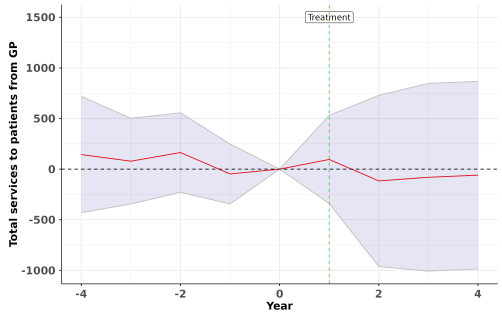
(b) Effect in region 1 - Montreal and neighbouring regions (Variation)



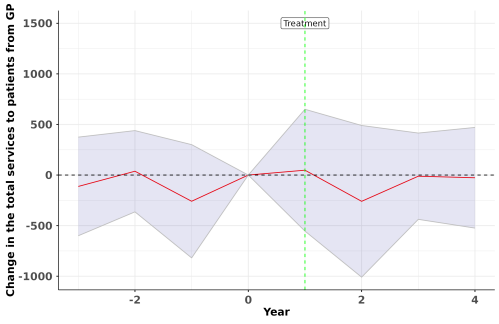
(c) Effect in region 2 - Quebec and neighbouring regions (Level)



(d) Effect in region 2 - Quebec and neighbouring regions (Variation)



(e) Effect in region 3 - Estrie (Level)



(f) Effect in region 3 - Estrie (Variation)

Figure B.10: Heterogenous effect on the total services given by SP per group of regions

C Proofs

C.1 Proof of Proposition 1

Consider the function:

$$Q^h(\theta, \mathbb{E}\mathbf{y}) = \sum_j \sum_i (\mathbb{E}y_{ij} - y_j^0)^2 F([\theta - \mathbb{E}y_{ij}]/h) + (y_j^1 - \mathbb{E}y_{ij})^2 F([\mathbb{E}y_{ij} - \theta]/h)),$$

but noting that $\mathbb{E}y_{ij}$ can only be either y_j^0 or y_j^1 under Assumption 1, we have:

$$Q^h(\theta, \mathbb{E}\mathbf{y}) = \sum_j n_j (y_j^1 - y_j^0)^2 F([\theta - y_j^1]/h) + (n - n_j)(y_j^1 - y_j^0)^2 F([y_j^0 - \theta]/h)),$$

or

$$Q^h(\theta, \mathbb{E}\mathbf{y}) = \sum_j (y_j^1 - y_j^0)^2 [n_j F([\theta - y_j^1]/h) + (n - n_j) F([y_j^0 - \theta]/h))] \geq 0,$$

Note that as $h \to 0$, Q^h converges to 0 iff $\theta \in (\max_j y_j^0, \min_j y_j^1)$. This implies that the global minimum of Q^h is necessarily such that $\theta \in (\max_j y_j^0, \min_j y_j^1)$ for h small enough. The global minimum therefore solves the first order condition.

Taking the first derivative, we have:

$$\frac{1}{h} \sum_j (y_j^1 - y_j^0)^2 [n_j f([\theta - y_j^1]/h) - (n - n_j) f([y_j^0 - \theta]/h)],$$

so the second derivative is:

$$\frac{1}{h^2} \sum_j (y_j^1 - y_j^0)^2 [n_j f'([\theta - y_j^1]/h) + (n - n_j) f'([y_j^0 - \theta]/h)].$$

For any $\theta \in (\max_j y_j^0, \min_j y_j^1)$, we have $f'([\theta - y_j^1]/h) > 0$ and $f'([y_j^0 - \theta]/h) > 0$. The function Q^h is therefore strictly convex over the identified set.

This shows that, for h small enough, Q^h has a unique minimum, which lies strictly inside the identified set.

Now, assume that $\bar{\mathbf{y}}$ converges in probability to $\mathbb{E}\mathbf{y}$. From the continuous mapping theorem, $Q^h(\theta, \bar{\mathbf{y}})$ converges pointwise in probability to $Q^h(\theta, \mathbb{E}\mathbf{y})$. But, because θ can be taken in a compact space (e.g. the identified set) and the first derivative of $Q^h(., \mathbb{E}\mathbf{y})$ is bounded, the convergence is uniform.

To summarize, $Q^h(., \mathbb{E}\mathbf{y})$ is continuous, uniquely minimized at $\theta = \theta_0$ (strictly inside the

identified set), θ can be taken in a compact set without loss, and $Q^h(.,\bar{\mathbf{y}})$ converges uniformly in probability to $Q^h(.,\mathbb{E}\mathbf{y})$. Thus, $\hat{\theta}$ converges in probability to $\theta_0 \in (\max_j y_j^0, \min_j y_j^1)$.

C.2 Proof of Proposition 2

From analysis of Section 6, we have:

$$\bar{c}_{j,t} = \psi_j + \frac{\bar{\gamma}_j}{1+\bar{\gamma}_j}\bar{c}_{j,t-1} + \bar{\varepsilon}_{j,t},$$

which is an ergodic process. As a result, we obtain:

$$\mathbb{E}c_{ij,t}^0 = \alpha_0 + \alpha_1 n_j,$$

and

$$\mathbb{E}c_{ij,t}^0 = \beta_0 + \beta_1 n_j.$$

Now, Assumption 1 imposes that $\mathbb{E}c_{ij,t}^0 < \mathbb{E}c_{ij,t}^1$ which is equivalent to the condition in Proposition 2.

C.3 Additional algebra for the structural model

C.3.1 SP's optimal decision

For the SP's problem, the Lagrange conditions are:

$$\frac{1}{\gamma_{ij}}(c_{ij,t} - a_{ij,t}) = \lambda,$$

where λ is the Lagrange multiplier, $\gamma_{ij} = \gamma_0(1-g_{ij}) + \gamma_1 g_{ij}$ if $i > 0$ and $\gamma_{0j} = 1$.

If the constraint is not binding, then $\lambda = 0$ and the solution is $c_{ij,t} = a_{ij,t}$. Otherwise, we easily obtain:

$$\lambda = \frac{T_{j,t} - a_{0j,t} - \bar{a}_{j,t}}{\bar{\gamma}_j + 1}.$$

C.3.2 Optimal decision: observable quantities

From the queuing process, we obtain:

$$a_{ij,t+1} = b_{ij,t} - [c_{ij,t}^* - a_{ij,t}],$$

thus

$$a_{ij,t+1} = b_{ij,t} - \frac{\gamma_{ij}}{\bar{\gamma}_j + 1}(T_{j,t} - a_{0j,t} - \bar{a}_{j,t}) \tag{9}$$

Summing (9) over i, we obtain:

$$\bar{a}_{j,t+1} = \bar{b}_{j,t} - \frac{\bar{\gamma}_j}{\bar{\gamma}_j + 1}(T_{j,t} - a_{0j,t} - \bar{a}_{j,t}) \tag{10}$$

We also know that (summing the optimal solution over $i > 0$):

$$\bar{c}^*_{j,t} = \frac{1}{\bar{\gamma}_j + 1}\bar{a}_{j,t} + \frac{\bar{\gamma}_j}{\bar{\gamma}_j + 1}(T_{j,t} - a_{0j,t}),$$

or equivalently:

$$\bar{a}_{j,t} = (\bar{\gamma}_j + 1)\bar{c}^*_{j,t} - \bar{\gamma}_j(T_{j,t} - a_{0j,t}), \tag{11}$$

and

$$\bar{a}_{j,t-1} = (\bar{\gamma}_j + 1)\bar{c}^*_{j,t-1} - \bar{\gamma}_j(T_{j,t-1} - a_{0j,t-1}),$$

so

$$\bar{a}_{j,t} - \bar{a}_{j,t-1} = (\bar{\gamma}_j + 1)(\bar{c}^*_{j,t} - \bar{c}^*_{j,t-1}) - \bar{\gamma}_j(T_{j,t} - a_{0j,t}) + \bar{\gamma}_j(T_{j,t-1} - a_{0j,t-1}). \tag{12}$$

Then, using (4) and a lagged version of (9), we obtain:

$$c^*_{ij,t} = b_{ij,t-1} - \frac{\gamma_{ij}}{\bar{\gamma}_j + 1}(T_{j,t-1} - a_{0j,t-1} - \bar{a}_{j,t-1}) + \frac{\gamma_{ij}}{\bar{\gamma}_j + 1}(T_{j,t} - a_{0j,t} - \bar{a}_{j,t}).$$

Finally, substituting with Equation (12), we obtain the expression in the text.

C.3.3 Cumulative treatment effect

Consider

$$\bar{c}_{j,t}(0) = \psi_j + \frac{\bar{\gamma}_j}{1 + \bar{\gamma}_j}\bar{c}_{j,t-1}(0) + \bar{\varepsilon}_{j,t},$$

and

$$\bar{c}_{j,t}(\delta) = \psi_j + \delta + \frac{\bar{\gamma}_j}{1 + \bar{\gamma}_j}\bar{c}_{j,t-1}(\delta) + \bar{\varepsilon}_{j,t},$$

where $\bar{c}_{j,t-1}(0) = \bar{c}_{j,t-1}(\delta)$. We easily obtain: $\bar{c}_{j,t}(\delta) - \bar{c}_{j,t}(0) = \delta$. We similarly have:

$$\bar{c}_{j,t+1}(\delta) - \bar{c}_{j,t+1}(0) = \delta + \frac{\bar{\gamma}_j}{1 + \bar{\gamma}_j}(\bar{c}_{j,t}(\delta) - \bar{c}_{j,t}(0)],$$

From which we get: $\bar{c}_{j,t+1}(\delta) - \bar{c}_{j,t+1}(0) = \delta + \delta\frac{\bar{\gamma}_j}{1+\bar{\gamma}_j}$. Continuing the sequence, we obtain:

$$\bar{c}^*_{j,t+\tau}(\delta) - \bar{c}^*_{j,t+\tau}(0) = \delta \sum_{s=0}^{\tau} \left(\frac{\bar{\gamma}_j}{1+\bar{\gamma}_j} \right)^s .$$

A simple variable change gives the result in the text: $T = t + \tau$ and $S = \tau - t$, which runs from $s = 0$, (or equivalently $\tau = t$), up to $s = \tau$ (or equivalently $\tau = T$).

C.3.4 Relative preference

We are interested in the following inequality:

$$\frac{c^1_{ij,t}}{a^1_{ij,t}} \geq \frac{c^0_{ij,t}}{a^0_{ij,t}}$$

Using the optimal solution (4), we can compute:

$$\frac{c_{ij,t}}{a_{ij,t}} = 1 + \frac{\gamma_{ij}}{a_{ij,t}} \frac{T_{j,t} - a_{0j,t} - \bar{a}_{j,t}}{\bar{\gamma}_j + 1}.$$

So, the inequality holds iff

$$\frac{\gamma_1}{\gamma_0} \leq \frac{a^1_{ij,t}}{a^0_{ij,t}}$$

And, using (9), it is equivalent to

$$\frac{b^0_{ij,t}}{b^1_{ij,t}} \leq \frac{\gamma_0}{\gamma_1}$$

Which is simply:

$$\frac{\alpha_0 + \alpha_1 n_j + \eta^0_{ij,t}}{\beta_0 + \beta_1 n_j + \eta^1_{ij,t}} \leq \frac{\gamma_0}{\gamma_1}$$

or

$$\gamma_1\alpha_0 + \gamma_1\alpha_1 n_j + \gamma_1\eta^0_{ij,t} \leq \gamma_0\beta_0 + \gamma_0\beta_1 n_j + \gamma_0\eta^1_{ij,t}$$