Designing Dynamic Reassignment Mechanisms: Evidence from GP Allocation*

Ingrid Huitfeldt[§] Victoria Marone[†] Daniel Waldinger[‡]

April, 2024

Please click here for updated version

Abstract

Many centralized assignment systems seek to not only provide good matches for participants' current needs, but also to accommodate changing preferences and circumstances. We study the problem of designing such a mechanism in the context of Norway's system for dynamically allocating patients to general practitioners (GPs). We provide direct evidence of misallocation under the current system—patients sitting on waitlists for each others' GPs, but who cannot trade—and propose an alternative mechanism that adapts the Top-Trading Cycles (TTC) algorithm to a dynamic environment. In contrast to the static case, dynamic TTC may leave some agents worse off relative to a status quo where trades are not permitted, introducing a new set of concerns about fairness. We then estimate a structural model of switching behavior and GP choice and empirically evaluate how this mechanism would perform relative to the status quo. While introducing TTC would on average reduce waiting times and increase patient welfare—with especially large benefits for female patients and recent movers—patients endowed with undesirable GPs would be harmed. Adjustments to the priority system can avoid harming this group while preserving most of the gains from TTC.

Keywords: top-trading cycles, dynamic allocation mechanism, waiting lists
JEL Codes: I18, D47, D04

*We are grateful to Bob Town, Jacob Leshno, Nikhil Agarwal, Parag Pathak, Irene Lo, Itai Ashlagi, Al Roth, Paulo Somaini, Liran Einav, and seminar participants at Imperial College London, Penn State, Duke, Minnesota, Boston College, ASU, Berkeley, Cornell, Chicago Harris, MIT Sloan, and the University of Pennsylvania for helpful comments and suggestions. Joridan Barash, Andrew Kim, and Junrui Lin provided outstanding research assistance. [§] Department of Economics, BI Norwegian Business School (email: ish@ssb.no) [†]Department of Economics, University of Texas at Austin (email: marone@utexas.edu) [‡]Department of Economics, New York University (email: danielwaldinger@nyu.edu)

I Introduction

Centralized (non-price) assignment mechanisms are used to allocate many important resources in the economy, including schools, jobs, housing, and healthcare. A rich theoretical and empirical literature has studied the design of such mechanisms. Much of this work has focused on providing good matches for participants' current needs. However, in many of these markets, agents' preferences over objects may change over time. Students may wish to transfer schools; public housing residents may want to down-/up-size as household composition changes; workers may want to relocate. Much less is known about how to design markets when there are repeated matching opportunities. Even in markets with sophisticated centralized assignment mechanisms, aftermarkets and reassignment systems are often less carefully designed.

This paper studies the problem of dynamically *re*-allocating agents to objects, when agents' preferences (or circumstances) change over time. We make three conceptual and empirical contributions. First, we provide direct empirical evidence of unrealized gains from trade in an important dynamic assignment market—the market for general practitioners (GPs) in Norway. Second, we propose an alternative mechanism that adapts a standard tool for centralized reassignment in a static setting—the Top-Trading Cycles (TTC) algorithm—to a dynamic environment, and clarify the incentive and distributional challenges that arise. Finally, we develop and estimate a model of patient preferences and choice over GPs in order to empirically evaluate counterfactual mechanisms. We evaluate both utilitarian and distributional welfare outcomes within a dynamic equilibrium model of a patient-GP (re)allocation system.

Our empirical setting is the Norwegian primary care system. As in many national health insurance schemes, every individual in Norway has a formally assigned GP who acts as a first point of contact and gatekeeper to secondary care. In principle, individuals (or hereafter, patients) have free choice of GP. In practice, however, each GP has a cap on the number of patients they can have on their "panel," thus creating capacity constraints that limit patients' effective choice of GP (Lovdata, 2012). In an effort to increase choice, in November 2016, Norway began allowing patients to join waitlists for oversubscribed GPs, while keeping their spot on their current GP's panel. Patients are permitted to stand on at most one waitlist at a time, and are assigned from the waitlist on a first-come, first-served basis to vacancies on the desired GP's panel.

The starting point for this paper is the observation that at any given time, there are are many patients waiting for each others' GPs. In December 2019, 15 percent of patients on a waitlist could be reassigned through a single run of the TTC algorithm. Women, young people, and patients who live further from their current GP were particularly likely to be among the patients who could have been immediately reassigned. A simple mechanical simulation suggests that if TTC had been run every month since the inception of waitlists—holding patients' GP choices fixed—the number of patients standing on waitlists would have been 23 percent lower at the end of 2019. Like in a static setting, allowing reassignments only if there is a vacancy substantially limits potential gains from trade.

Despite clear evidence of unrealized gains from trade, the consequences of incorporating TTC into Norway's GP reassignment system are not obvious. A general challenge in designing reassignment mechanisms is that market participants may face dynamic incentives. Their choices affect not only what they receive today, but also future matching opportunities. Many of the strong theoretical results established in static settings—where TTC is known to be both efficient and strategy-proof (Shapley and Scarf, 1974; Roth, 1982)—do not necessarily apply in dynamic settings. As a result, the theoretical literature has not provided general characterizations of optimal dynamic matching mechanisms.

Thus, while introducing TTC to implement cycles of switches seems like a clear choice, there are reasons for pause. First, patients' equilibrium responses to changes in the assignment mechanism may offset the mechanical reductions in waiting time that arise when patients' actions are held fixed. Further, introducing TTC would not necessarily be a Pareto Improvement. While average waiting times fall substantially in the mechanical simulation, some patients are harmed—they experience *longer* waiting times because the panel slot they would have taken under the current system is instead given to another patient who arrived later, but who can form a cycle with the owner of that slot. These dynamic effects represent a more subtle aspect of TTC: only a patient with an oversubscribed GP (whose endowment is a scarce resource) has the opportunity to "trade." Patients with undersubscribed GPs are effectively de-prioritized, and may experience systematically longer waiting times. To the extent that perceived unfairness is undesirable, the overall gains from implementing TTC may not in fact appear worthwhile. The size of both the overall gains and the associated distributional consequences are ultimately an empirical question.

The rest of the paper studies the equilibrium impacts of introducing TTC and other related matching algorithms to Norway's existing waitlist system. We address two key empirical challenges. First, patients' GP choices will likely respond to changes in waitlist lengths as well as to changes in their beliefs about how quickly waitlists will move, both of which may change across mechanisms. Second, the number of patients standing on waitlists grew rapidly over our sample period (2016–2019), meaning the data does not contain direct information on what a stationary environment would look like. Addressing these challenges requires economic modelling on two fronts. We first formulate a demand model of patients' decisions to switch GP, and their GP choice conditional on switching. We then introduce a dynamic equilibrium model of a patient-GP reassignment system, which will allow us to predict outcomes in a stationary equilibrium.

Our demand model features exogenous (in)attention and endogenous GP choice. In a given period, patients stochastically "pay attention" with probability that depends on both time-varying (e.g., a recent move) and time-invariant (e.g., demographic) factors. Attentive patients then make a discrete choice over GPs, choosing whether to switch to a GP with available panel slots, join a GP's waitlist, or remain with their current GP. Importantly, a patient's choice depends not only on which GP is most preferred if she could be reassigned immediately, but also on her relative (dis)satisfaction with her current GP, which she would retain while waiting on a waitlist.

We estimate the model parameters via a Gibbs' sampler using monthly administrative data on Norway's GP assignment system. Our estimates imply substantial horizontal differentiation across GPs, suggesting large returns to an efficiently designed assignment system. Much of this differentiation is driven by geographic location, but we also estimate that patients have strong preferences for a doctor of the same gender (worth the equivalent of 6–7 minutes of travel time) and similar age (1 minute). The estimated attention probabilities closely match switch request rates by age, gender, and whether and how far the patient recently moved. Finally, our preferred specification estimates an annual discount factor of approximately 0.91, consistent with reduced form evidence that patients' GP choices are responsive to waitlist lengths.

We apply these estimates within a dynamic equilibrium model of Norway's patient-GP reassignment system. Our model is calibrated to match the basic elements of the Norwegian setting, including GP characteristics and the rate at which patients age, die, and move between municipalities. We define an equilibrium in which there is a fixed point between patients' beliefs about waiting time and their optimal GP decisions, where beliefs match the long-run stationary distributions generated by optimal behavior. Our simulations imply that under Norway's status quo mechanism, 9 percent of the population would be standing on a waitlist in the stationary equilibrium, and the average patient could expect to wait over a year to switch GP.[1]

[1] As of February 2024, 6.5 percent of the population is standing on a GP waitlist, more than twice the number at the end of our sample period. Thus, our prediction that the waitlist would continue to grow dramatically

Our primary counterfactual of interest runs the TTC algorithm at the end of each month on top of Norway's current waitlist system, after all naturally arising vacancies have been filled from waitlists. Relative to the status quo, the gains from introducing TTC are equivalent to reducing every patient's travel time to their GP by 0.7 minutes. Much of this improvement is due to patients obtaining GPs who are better matched to them in terms of location (0.4 minutes) and gender. Overall, TTC benefits the majority of patients, and in particular younger and female patients, who are most likely to request to switch GPs and to use waitlists.

Some patients, however, are harmed by the introduction of TTC. In particular, patients whose GPs are undersubscribed face longer waiting times, and are overall worse off. This harm is driven both directly by the fact that TTC prioritizes patients with desirable endowments, and indirectly by these patients' increased willingness to wait for the most desirable GPs. Since it may seem unfair to disadvantage patients who already have less desirable GPs, we consider two alternative mechanisms intended to mitigate these distributional consequences. First, and most simply, we implement the patient-proposing deferred acceptance (DA) algorithm each month instead of TTC. This mechanism strictly respects first-come first-served waiting time priority, and thus does not make *any* patients worse off relative to the status quo.[2] However, it produces almost negligible gains relative to the status quo, illustrating a fundamental trade-off between avoiding envy and exploiting gains from trade. Second, we implement a "TTC with priority" (TTCP) algorithm that prioritizes patients with undersubscribed GPs for panel vacancies.[3] The results are encouraging. Although mean utilitarian welfare is lower than under TTC, TTCP still improves outcomes relative to the status quo for patients with both over- and undersubscribed GPs.

Related Literature. First and foremost, this paper contributes to a growing empirical literature on centralized matching markets. Recent work has studied numerous applications using static models, particularly in the context of school choice (Abdulkadiroğlu, Agarwal and Pathak, 2017; Agarwal and Somaini, 2018; Calsamiglia, Fu and Güell, 2020; Fack, Grenet and He, 2019; Kapor, Neilson and Zimmerman, 2020) and specialized labor markets (Agarwal, 2015). There has also been recent work on dynamic assignment problems, in which agents and vacant objects arrive to the mechanism stochastically over time (Waldinger, 2021; Agarwal et al., 2021; Verdier and Reeling, 2021; Lee, Ferdowsian and Yap, 2024). Our paper belongs

is at least qualitatively consistent with what has occurred.

[2]DA is distinct from Norway's status quo mechanism because DA allows trades among patients at the top of each waitlist, whereas Norway's system requires a vacancy before any reassignments can be made.

[3]This proposal is analogous to the priority given to blood type O patients for blood type O donors in organ allocation. We are grateful to Al Roth and Itai Ashlagi for this suggestion.

to a distinct class of problems in which agents and vacant objects still arrive stochastically, but in addition, agents bring an endowed object with them when they arrive. The role of endowments is also central to the analysis of Narita (2018) and Kapor, Karnani and Neilson (2024), which study aftermarkets in centralized school assignment. Larroucau and Ríos (2022) estimate a dynamic model of college major choice in which students can be reassigned, focusing on learning and congestion externalities. Combe, Tercieux and Terrier (2022) study the reassignment system for French teachers assuming teachers truthfully report their (static) preferences. The present paper provides novel empirical evidence on the design of dynamic reassignment mechanisms when agents are forward-looking and waiting time is the primary rationing mechanism.

To our knowledge, there is no theoretical characterization of optimal mechanisms for the class of models we consider, motivating an empirical approach to evaluating alternative market designs. Nonetheless, our proposed mechanisms draw from an extensive theoretical matching literature. The TTC algorithm was proposed by Shapley and Scarf (1974) and is attributed to David Gale. Modified to respect endowments, it is known to be efficient, strategy-proof (Roth, 1982), and individually rational (Abdulkadiroğlu and Sönmez, 1999) in a static environment. These properties motivate our proposal to adapt TTC to a dynamic environment, but as we demonstrate, may break down when participants face dynamic incentives. The trade-off for market participants between shorter waiting times and a preferred assignment is central to theoretical work on dynamic assignment (Su and Zenios, 2004; Bloch and Cantala, 2017; Arnosti and Shi, 2020; Leshno, 2022; Che and Tercieux, 2023). While many of these idea remain relevant in dynamic *re*-assignment, the direct connection between supply and demand that arises when agents bring endowments creates a distinct set of theoretical challenges. Work that considers re-assignment largely abstracts away from the dynamic incentive issues considered here (Ashlagi, Nikzad and Strack, 2022; Akbarpour, Li and Gharan, 2020; Akbarpour et al., 2023; Combe, Tercieux and Terrier, 2022), or focuses on a limited form of agent heterogeneity (Baccara, Lee and Yariv, 2020). Narita (2018) and Feigenbaum et al. (2020) both propose strategy-proof mechanisms for a re-matching model with general preferences, but limit consideration to two-periods.

Our paper also contributes to the literature studying patient preferences over health care providers. While a long literature has estimated demand for hospitals and other facilities (e.g., Tay, 2003; Cutler, Huckman and Landrum, 2004; Ho, 2006; Pope, 2009), a smaller literature has focused on demand for individual physicians, and these papers have focused primarily on heart surgeons (e.g. Wang et al., 2011; Kolstad, 2013). Few papers have modelled individuals' demand for primary care providers, perhaps due to lack of good data in such settings. Existing work has studied the role of gender homophily (Godager, 2009) and of information on quality (Santos, Gravelle and Propper, 2017; Bensnes and Huitfeldt, 2021; Empel, Gravelle and Santos, 2023; Brown et al., 2023). Our setting allows us to evaluate the extent to which patients are willing to wait for a desired GP. Beyond patient demand, several papers have established the importance of primary care for health outcomes (Bailey and Goodman-Bacon, 2015; Baker, Bundorf and Beeson Royalty, 2019; Fadlon and Van Parys, 2020; Ding et al., 2021; Dahlstrand, 2022; Mora-García, Pesec and Prado, 2023). We view our demand estimates as embedding patients' perceptions about any health benefits they would enjoy from different GPs. In focusing on patients' revealed preferences for GPs, rather than on health outcomes, we are similar to Brown et al. (2023) and Chartock (2021), though our focus is on dynamic incentives rather than information frictions.

Finally, as in other countries, the existence of a national health insurance scheme motivates the need for a non-price based mechanism to allocate healthcare resources at the point of service. Our paper contributes to a literature studying the optimal design of such mechanisms. Early work focused on whether limiting capacity and running waitlists for non-emergency services could deter low-value care (Nichols, Smolensky and Tideman, 1971; Propper, 1990, 1995; Gravelle and Siciliani, 2008*b*), and subsequent work has studied the optimal design of prioritization schemes on such waitlists (Gravelle and Siciliani, 2008*a*; Gruber, Hoe and Stoye, 2023; Shen et al., 2020). Our work is among the first papers to bring the tools of market design to bear on this topic. In a closely related setting, Mark (2021) studies the Canadian primary care system, in which the reallocation mechanism is decentralized, and patients who wish to switch GPs must exert costly effort to find vacancies. In contrast, we study the design of a centralized mechanism.

The paper proceeds as follows. Section II.A introduces our setting and data and demonstrates the existence of gains from trade among waiting patients. Section III presents the structural model of patient attention and GP choice. Section IV describes our estimation procedure and results. Section V presents our counterfactual simulations, and Section VI concludes.

II Background, Data, and Descriptive Evidence

II.A Empirical Setting

Norway has a comprehensive national health insurance scheme primarily financed by general taxation. Patient cost-sharing at the point of service is nonzero for most outpatient care, but is still limited. Healthcare utilization is primarily managed via supply-side forces. Central among these forces is a gate-keeping system whereby patients need a referral from a primary care provider before receiving specialist care. Such providers therefore play a central role in the healthcare system.

Primary care is almost exclusively provided by GPs.[4] GP care is organized via a *patient panel* system, whereby every person enrolled in the national health insurance scheme is assigned to a specific GP. Patients enrolled on a given panel are in general only permitted to visit that GP for their primary care needs.[5] Similar primary care systems exist in Canada, Great Britain, Italy, and Sweden (among others), as well as in the context of Health Maintenance Organizations (HMOs) in the US. One reason for maintaining a centralized administrative linkage between a patient and a specific GP is to allow for a "capitated" payment model, under which GPs receive a fixed payment for each person enrolled on their patient panel. Norway uses a partially capitated payment model, meaning GPs receive a fraction of their revenue from capitated payments (on average 30 percent) and the remainder from fee-for-service payments.

The supply of GPs is regulated through a fixed number of government contracts.[6] Similar to Medicare and Medicaid physician contracts in the US, government contracts involve take-it-or-leave-it payment terms, with some exceptions made to attract physicians to rural areas. One important difference in Norway is that at the time a GP enters into a government contract, both parties must agree on a maximum number of patients that the GP's panel can take.[7] In

[4]As of 2023, nurses in Norway do not have prescribing or referring authority outside of a few special cases. While it is common for nurses to handle well visits for children and adolescents, adult patients must typically consult a GP for the majority of their primary care needs (Robstad et al., 2022; Hansen, Boman and Fagerström, 2020).

[5]There are some exceptions; for example, patients have the right to seek a second opinion from another GP on a matter already discussed with their own GP. Primary care needs could include periodic well-visits, non-urgent sick visits, obtaining prescriptions or referrals to specialist care, or receiving documentation for sick leave through the national sick leave scheme.

[6]While GPs can also practice outside of the national health insurance scheme, the market for private practice primary care remains small or non-existent in most of the country (The European Observatory On Health Systems and Policies, 2023)

[7]Operationally, it is local municipal governments that work with GPs to set panel caps and negotiate any

entering into the contract, the GP agrees to take on a workload sufficient to serve all patients that enroll on their panel, up to the agreed panel cap. GPs are required to be able to provide an appointment to patients within 5 working days, which in part motivates the existence of panel size caps (Lovdata, 2012).

Patients make GP enrollment elections via a nationally centralized online platform.[8] The system operates on a rolling basis, without any special enrollment periods or forced re-enrollment decisions. Enrollment changes take effect on the first day of the next month, and GPs have no way to control which patients enroll on their panel. When a GP retires or quits, patients receive six months' notice and can either switch GP or remain on the panel of the replacement GP.[9] Newborns are by default automatically assigned to their mother's GP, regardless of whether the panel cap is violated. Thus, every patient is assigned to a GP at all times.

While patients in principle have free choice over GPs, in practice the panel caps generate capacity constraints. A GP with no open slots on their panel is listed as "unavailable" on the online enrollment platform. Patients that wish to switch GP immediately may therefore only choose among the set of GPs with open slots on their panels (henceforth, "open panels"). Prior to 2016, a patient would simply need to check back later if their desired GP was unavailable.[10] In November 2016, a new functionality was introduced whereby patients could join a waitlist. Patients are permitted to join only one waitlist at a time, but can switch which waitlist they are on (if any) as much as they would like.[11] Panel slots that become available are filled from the waitlist on a first-come, first-served basis, according to the point in time at which each patient joined that waitlist. Once a patient joins a waitlist, they commit to being reassigned to the target GP once they reach the front of the list; there is no opportunity to renege except to remove oneself from the waitlist before it is too late.

---

other idiosyncratic features of a GP's contract, such as reduced working hours or coverage for parental leave spells.

[8]The online platform is publicly available at https://tjenester.helsenorge.no/bytte-fastlege. Appendix Figure A.1 provides a screenshot of the interface. Patients are permitted to switch GPs at will up to two times per year (though this constraint rarely binds), with additional switches granted for qualifying life events.

[9]If the replacement GP has a lower panel cap, a randomly selected subset of patients are administratively reassigned, without their explicit consent, to another GP in the local area.

[10]The "check back later" mechanism is still used in many GP allocation systems, such as Canada, the UK, Italy, Sweden, Denmark, and as far as the authors are aware, all HMOs in the US.

[11]The centralized web platform provides information on the number of patients on each waitlist at any given time (c.f. Appendix Figure A.1), and if a patient is logged in, their personal position on the waitlist they are standing on.

II.B Data

Our data are derived from two main sources. First, we observe detailed administrative data on the GP assignment system itself (*Fastlegedatabasen*). These data include a full history of patient enrollment, waitlist spells, and GP characteristics. We can therefore reconstruct the state of each GP's panel and waitlist at any point in time. Second, we link these data to register data from Statistics Norway on individual demographics, including age, gender, family relationships, income, education, and monthly municipality of residence. Further details are provided in Appendix A.1. We have data for the period 2014–2019.

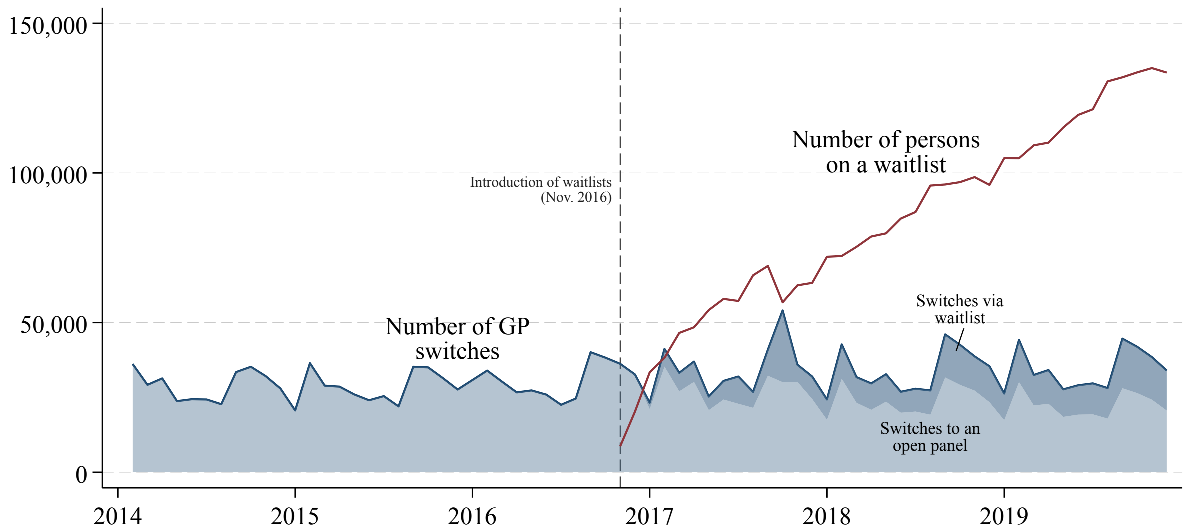
Figure 1 shows the number of GP switches and waitlist use over time. GP switches take one of two forms: (i) standard switches to an open panel, and (ii) switches that occur once an individual reaches the front of a waitlist and is reassigned. Between 2014 and 2019, there were an average of 5,259,076 patients per month in the GP allocation system. An average of 34,100 patients (0.6 percent) switched their GP each month.[12] Once waitlists were introduced in November 2016, an average of 28 percent of switches were executed via a waitlist (for an otherwise full GP panel), whereas the remaining 72 percent were still switches to an open GP panel. The waitlists have grown steadily since their introduction. By the end of 2019, 133,538 people were standing on a waitlist (2.5 percent of all individuals in the system). As of August 2023, this number had risen to 328,506 (6.2 percent of all individuals).

Patients. Our primary restriction on patients is to exclude those who are under 16 years old. Children's interactions with the healthcare system, including GP enrollment, are formally managed by parents until a child turns 16. There are also special exceptions given to children that allow them to bypass GP panel caps in certain situations, as well as to switch GPs alongside a parent without themselves going through a waitlist. As these factors would complicate our analysis to a substantial degree, we limit our focus to the experience of adults.

Table 1 provides summary statistics on patients, focusing on the three-year period from 2017 to 2019. The sample is an unbalanced panel at the patient-month level (panel entry occurs at age 16 or immigration; panel exit occurs at death or emigration). The first column describes the full sample. There are 4.78 million unique patients, representing the universe of over-16 individuals registered as resident in Norway and covered under the national insurance scheme over this period. Patients on average are 47 years old and have income of 408,374 NOK. Just over 7 percent of individuals are temporary residents, 32 percent have post-secondary

[12]This statistic captures only voluntary switches. An additional 0.6 percent of patients were on average administratively reassigned each month due to GP panel downsizing, exit, or other reasons.

Figure 1. Number of GP Switches and Waitlist Use Over Time



Notes: The figure shows the number of GP switches per month and the stock of individuals standing on a waitlist each month between 2014 and 2019, including both adults and children. GP switches are decomposed into standard switches to an open GP panel (light blue) and switches that occur only after going through a waitlist (dark blue).

education, and 10 percent moved to a new municipality at least once during 2017–2019.

In terms of GP choice, 19 percent of patients requested to switch their GP at least once (4 percent of patients did so more than once). The average travel time (by car) to a patient's GP was 10.7 minutes; the median was 5.8 minutes.[13] More than half (58 percent) of patients have a GP of the same gender as themselves. In terms of waitlist use, 9 percent of individuals were at any point on a waitlist over this period. Conditional on ever being on a waitlist, the average number of months on a waitlist was 6.5.

The remaining three columns of Table 1 break up the full patient sample into three subgroups: (i) patients who never switched their GP nor joined a waitlist, (ii) patients who switched GP but never joined a waitlist, and (iii) patients who joined a waitlist (and may or may not have successfully switched their GP). Several patterns are apparent. First, gender homophily appears to be a driver of switching behavior. While 56 percent of patients on waitlists currently held a GP with the same gender as themselves, 64 percent of them were waiting for a GP of the same gender. Waiters are also much more likely to be female, which could be driven by a scarcity of female GPs, or a stronger gender homophily preference among female patients.

[13]Travel time is measured between the population-weighted centroid of patients' municipality of residence and the address of their GP's office.

Table 1. Patient Summary Statistics

Sample demographic	Full Sample	Never used waitlist: Never switched	Never used waitlist: Ever switched	Ever used waitlist
Number of individuals	4,780,647	3,875,753	497,111	407,783
Pct. of individuals		0.81	0.10	0.09
Demographics				
Pct. female	0.50	0.48	0.51	0.64
Age	47	49	40	42
Pct. with post-secondary education	0.32	0.31	0.32	0.37
Annual income (000 NOK)	425	437	358	390
Pct. temporary resident	0.07	0.07	0.08	0.08
Pct. rural	0.30	0.31	0.30	0.28
Pct. ever moved	0.10	0.06	0.34	0.26
Choice of GP				
Pct. ever switched to open GP	0.13	—	1.00	0.28
Travel time to current GP (min.)	10.7	9.5	17.1	14.3
Pct. with GP of same gender	0.58	0.57	0.59	0.56
Use of waitlists				
Pct. ever on a waitlist	0.09	—	—	1.00
Number of months on a waitlist $\mid > 0$	6.5			6.5
Pct. waiting for GP of same gender	0.64			0.64
Travel time to wl. GP – curr. GP (min.)	-6.8			-6.8

Notes: The table provides summary statistics on adult patients present in the data from 2017–2019. Except where otherwise specified, all values in the table represent means over patient-months. "Ever" means at any point during 2017–2019. Moves are counted only if they are across municipalities. Age, gender, education, and income data are not available for temporary residents, so those means are only among permanent residents.

Second, there is a striking pattern with respect to moves. Among patients who never requested to switch their GP, only 6 percent of individuals moved municipality during this period. Among those who switched GP but never used a waitlist, 34 percent of individuals moved. Among those who used a waitlist, 26 percent moved. The high move rate among patients requesting to switch GP suggests that geographic proximity is an important factor in GP choice. The even higher move rate among patients who only open-switched is consistent with patients who are far from their current GP being less selective, i.e., more likely to simply settle with an open GP. A final observation on GP proximity is the last row of the table, which shows the difference in travel time between the GP a patient is on a waitlist for and the patient's current GP. Among patients who used a waitlist, the waitlist GP was on average 6.8 minutes closer than their current GP (on a base of 14.3 minutes travel time to current GP among patients in this group).

GPs. Table 2 provides summary statistics on the 6,470 GP panels active during 2017 to 2019. As with patients, the data consist of an unbalanced panel at the GP panel-month level.[14] Note that there is an important distinction between a *GP panel* and a *GP*. A GP panel is the administrative unit to which patients are actually linked. Each panel is then served by a GP (a licensed medical doctor). The GP that serves a given panel may change over time, independently of the patients enrolled on that panel.

The first column of Table 2 describes the full set of GP panels. The average panel had a cap of 1,138 and had available slots for 36 percent of months. Across all months between 2017 and 2019, the average GP serving these panels was 47 years old, 42 percent of GPs were female, and 11 percent were temporary GPs.[15] The average GP panel-month had 18 people standing on its waitlist, and the average enrollment-to-cap ratio was 94 percent.

Table 2. GP Panel Summary Statistics

	All	Undersubscribed	Oversubscribed
Number of GP panels	6,470	3,532	2,938
Pct. of GP panels	1.00	0.55	0.45
Panel characteristics			
Enrollment cap	1,138	1,143	1,135
Pct. months with available slots	0.36	0.70	0.06
Pct. rural	0.37	0.44	0.30
GP demographics			
Age	47	47	48
Pct. months with female GP	0.42	0.34	0.50
Pct. months with temporary GP	0.11	0.15	0.08
Panel enrollment stats.			
Num. enrollees on a waitlist	18	21	14
Num. waiting on waitlist	18	4	30
Num. enrollees / cap	0.94	0.87	1.00

Notes: The table provides summary statistics on the set of GP panels present in the data from 2017–2019. Except where otherwise specified, all values in the table represent means over GP panel-months. Oversubscribed GP panels are those which are at capacity for more than 75 percent of months. Enrollment and waitlist use statistics reflect the full population (including children under 16).

The remaining two columns of Table 2 separate GP panels into two subgroups: (i) those that were consistently undersubscribed, and (ii) those that were consistently oversubscribed. We define "consistently oversubscribed" to mean that a given GP panel was at capacity for more than 75 percent of the months that it appeared in the data from 2017–2019. There are several notable patterns. First, GP panels in urban areas and those served by female

[14] The average GP panel appears in the data for 28 months (out of 36 possible). At a given time, between 4,840 and 5,106 panels are in operation.

[15] Temporary GPs (*vikar*) are used while the primary GP (*fastlege*) is on parental leave or during times when the position is vacant and the municipality is actively searching for a new permanent GP.

GPs are substantially more likely to be oversubscribed. GP panels served frequently by temporary GPs are more likely to be undersubscribed. Second, among the current enrollees of undersubscribed panels, an average of 21 patients (2.1 percent of enrollees) were themselves standing on a waitlist (and therefore trying to *leave* the panel). Among oversubscribed panels, this is true for fewer patients—an average of 14 people per panel (1.2 percent of enrollees). Thus, while some GPs are systematically more demanded than others, there are still many patients requesting to switch away from over-demanded GPs. Finally, capacity utilization is high; even for undersubscribed GPs, most panel slots (87 percent) were occupied.

II.C Prevalence of Double Coincidence of Wants

This section provides direct evidence of unrealized gains from trade among patients on Norway's GP waitlists. This prima facie evidence motivates the structural model and counterfactual simulations we develop in the remainder of the paper.

Static Gains From Trade. The first-come, first-served waitlists used in Norway's GP allocation system do not permit trades among waiting individuals. Even when two people are each at the front of the waitlist for the other's GP, this "trade" cannot occur until there is a vacancy on one of the two panels, allowing one waiting patient to vacate their spot for the other. Note that the same problem also exists in a mechanism without formal waitlists. Two individuals could simultaneously check to see if there was an opening at one another's GP, see that there was not, and not be able to trade. The advantage (from the analyst's perspective) of the waitlist system is that we have a record of these preferences. In particular, we are able to observe whether there are individuals who want to trade with one another. Such trades could be bilateral between only two individuals, or multilateral among an arbitrary number of individuals.

Using the waitlists data, we begin by calculating the extent to which such "double coincidence of wants" exist. We search for possible trades using the well-established Top Trading Cycles algorithm (Shapley and Scarf, 1974; Abdulkadiroğlu and Sönmez, 1999), described formally in Section V.B. We first run the algorithm on the waitlists from the last month of our data, December 2019, using our full sample. At the end of this month, there were 133,332 individuals standing on a waitlist. These waiters were currently enrolled on almost every existing GP panel (4,963 out of 5,010), but were standing on the waitlist for only 3,695 unique

GPs.[16]

The Top Trading Cycles algorithm takes two primary sets of inputs: (i) the set of participating agents and their rank-order preference lists over objects, and (ii) the set of participating objects and their rank-order priority lists over agents. In our setting, agents (patients) have preference lists of length at most two. Patients standing on a waitlist first prefer their waitlist GP, and then their current GP. Patients not standing on a waitlist prefer only their current GP. Objects (GPs) have longer priority lists. They first prioritize all patients currently on their panel; this guarantees each patient an assignment no worse than their current GP. GPs then prioritize the patients on their waitlist in descending order of waiting time.[17] We run the algorithm and find that 20,377 people (15 percent of waiters) could have been immediately reassigned via TTC.[18] Appendix Table A.1 describes the types of adult patients that could be immediately reassigned. Reassigned patients tended to have a larger difference in travel time between their current and waitlist GPs, consistent with location preference heterogeneity being a key driver of gains from trade.

Sources of Gains from Trade. The fact that gains from trade exist tells us that patients must differ—at least to some extent—in their preferred GP. We explore the sources of this horizontal preference heterogeneity descriptively using a conditional logistic regression predicting which GP a patient chooses, conditional on making a switch request. Switch requests are either an immediate switch to an open GP panel, or a waitlist join. While there is some vertical GP differentiation (in the form of GP fixed-effects), the role of horizontal differentiation is substantially larger in explaining choices. The most important source of horizontal preference heterogeneity is patients' geographic location relative to GP offices. Disutility derived from travel time accounts for 87 percent of the total variance in patients' (observable) valuations of GPs, while GP fixed effects account for only 3 percent. Beyond travel time, gender and age homophily are important determinants of horizontal preference heterogeneity. Compared to male patients, female patients are willing to travel further to see a female GP. And compared to patients over age 45, patients under age 45 are willing to travel further to see a GP who is

[16] Appendix Figure A.2 provides a histogram of waitlist lengths across these 3,695 waitlists.

[17] There can be many allocations in the core. Appendix B.1 explains the specific algorithm we use to find and clear cycles, which selects a particular core allocation, as well as other implementation details. We do not explore the choice of cycle-clearing rules due to its computational complexity, but view it as an interesting direction for future work.

[18] There is substantial geographic heterogeneity in the fraction of waiters that could be reassigned. In rural Åsnes municipality, it is only 2 percent (out of 420 waiters, where waiters represent 6 percent of the population). In urban college town Bergen, it is 26 percent (out of 7,991 waiters, where waiters represent 3 percent of the population).

also under age 45. Appendix B.2 provides a detailed description of these results.

While this analysis is useful as a descriptive exercise, there are two reasons why it is not appropriate for our counterfactual exercises. First, choice patterns in our data are inconsistent with there being an explicit "waiting cost" that enters payoffs directly (we will discuss this point further in Section III). The "cost" of waiting is rather the opportunity cost of remaining with one's current GP. Second, a patient's choice of GP should therefore also depend on their value of their current GP, which this specification cannot explicitly account for. These considerations will motivate the demand model in Section III.

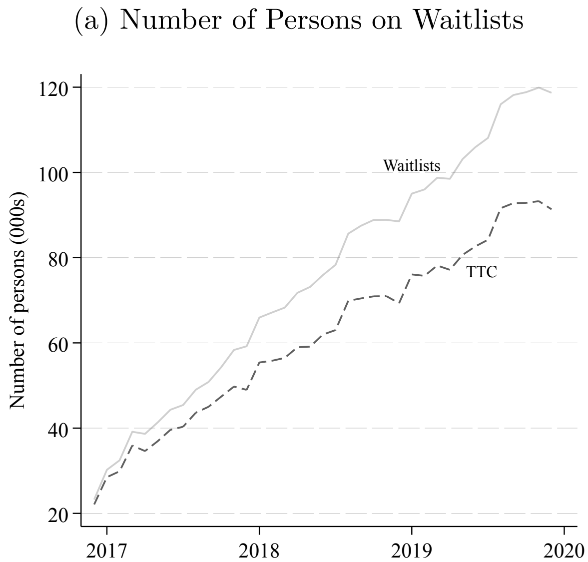
Impact on Evolution of Waitlists. The static analysis suggests that the gains from trade among patients with oversubscribed GPs may be substantial. The welfare impact of these trades, however, will depend on the extent to which they affect patients' waiting times and GP assignments in a dynamic environment. We can begin to get a sense of these dynamic effects by evaluating how the waitlist would have evolved if trades were processed periodically in the first three years the waitlists were operational. From November 2016 through December 2019, we run the TTC algorithm on the remaining waitlists at the end of each month. Importantly, patients' actions—GP switch requests—are held fixed. We then compare the number of reassignments and waiting times to those under Norway's current FCFS waitlist mechanism.[19]

Figure 2 presents the results of this exercise, comparing the status quo waitlists mechanism (Waitlists) to the counterfactual monthly implementation of Top-Trading Cycles (TTC). Panel (a) shows the number of people on waitlists each month. By the end of 2019, there would have been 23 percent fewer waiting patients under TTC. Panel (b) shows the realized waiting times. The blue series shows the average elapsed waiting time among the stock of individuals standing on waitlists each month, while the red series show the average waiting time among the flow of individuals who successfully get off a waitlist each month. Among reassigned patients, average waiting times would have been 29 percent shorter under TTC, suggesting that trades generated by TTC may lead to significant reductions in waiting times.

While TTC reduces average waiting time, there is wide variation in impacts across patients. Moreover, in contrast to a static environment, running TTC in a dynamic environment does not offer a Pareto improvement relative to a status quo in which it is not run. Some patients

[19] We compare simulations to simulations: simulated outcomes under FCFS waitlists to simulated outcomes under FCFS waitlists plus TTC. Appendix B.1 provides additional details about the implementation of this analysis.

Figure 2. Results of Running TTC on Historical Data



Notes: The figure shows the outcomes of running TTC at a monthly level on the historical waitlists data, holding all patient actions fixed. Panel (a) shows the number of persons standing on waitlists each month. Panel (b) shows the average elapsed waiting time among the sock of individuals standing on waitlists each month (blue) as well as the average waiting time among the flow of individuals who successfully got off a waitlist each month (red).

are harmed in the form of longer waiting times. This can happen because a slot that would have been taken by the first person on the waitlist might be filled earlier by a different patient who is further back on the waitlist but participated in a cycle with the patient previously in that slot. Appendix B.3 provides stylized examples of this phenomenon in our setting. Appendix Figure A.3 shows the distribution of waiting time differences among patients in our simulation. A minority of patients (4.5 percent) have longer waiting times under TTC because they are effectively de-prioritized relative to the status quo.[20]

Equilibrium Implications. The above analysis provides direct evidence of potential gains from trade if TTC were introduced into Norway's GP allocation system, but it has two important limitations. First, the simulation holds patient behavior fixed, even as waitlist lengths change. If the implementation of TTC changed the distribution of waitlist lengths, there is

[20]This point is related a broader debate about TTC in the market design literature. Because TTC finds gains from trade by relaxing priority rules (Abdulkadiroğlu and Sönmez, 2003), it yields an efficient allocation, but not a stable one. Further, allocations produced by TTC are not Pareto improvements over those produced by a Deferred-Acceptance algorithm (DA). In our setting, waiting time plays the role of priority, and Norway's status quo waitlists mechanism can be interpreted as a modified version of DA that only processes w-chains in the sense of Roth, Sönmez and Ünver (2004). TTC changes this priority structure, benefiting some patients at the expense of others.

good reason to think patients might have requested different GPs. Indeed, our analysis in Appendix B.2 shows that patients' choice of GP is responsive to waitlist length. Moreover, if patients understand that TTC may allow them to be reassigned faster, expected waiting times would fall even conditional on waitlist lengths, again potentially influencing patient choices. On one hand, these behavioral responses could improve the long-run efficiency of patient-GP matches, if patients became more willing to wait for their most preferred GP. On the other hand, it could exacerbate the distributional consequences seen above. Since patients enrolled with undersubscribed GPs will not benefit from the possibility of participating in a cycle, they may face longer waiting times as more patients queue for the most desirable GPs. The primary role of our structural model will be to assess the trade-offs between shorter average waiting times and better matches on the one hand, and the uneven distributional consequences generated by patients' behavioral responses on the other.

A second limitation of this analysis is that the market was far from a steady state during our sample period. Use of waitlists rose rapidly after their introduction, and has continued to do so after the end of our sample.[21] The long-run stationary distribution of Norway's GP allocation system—and the associated impact of introducing TTC—may be different when queues are systematically longer. Our counterfactual simulations will rely on a stationary demographic evolution and GP switching process, allowing us to evaluate the impacts of TTC in a long-run equilibrium.

III Model of GP Preferences and Choice

Section III.A presents a model of attention and GP choice that will form the basis of our estimation strategy. The model has two parts. The first is a model of limited attention in which patients stochastically consider switching GPs. The second is a model of GP choice in which attentive patients decide which GP to request, if any. Section III.B then presents a belief model under which patients map waitlist lengths into beliefs about waiting time.

[21] In August 2023, there were 330,000 patients on a GP waitlist, more than twice the number when our data end in December 2019.

III.A Attention, Preferences, and Choice

Patients are indexed by i, GPs by j, and time by t. Time is continuous. We model patient preferences at the time they consider switching GPs. At time t, patient i's preferences are represented by indirect flow utilities from being assigned to each GP $\mathbf{v}_{it} \equiv (v_{i1t}, ..., v_{iJt}) \in \mathbb{R}^J$ and a discount rate ρ. The patient has observable attributes X_{it} and is currently assigned to GP j_{0t}. We will suppress t subscripts when they do not affect the exposition.

A patient "pays attention" and considers switching GPs at Poisson rate p^{λ}_{it}.[22] When a patient is attentive, two things happen: (i) she draws new preferences $\mathbf{v}_{it} \sim F(\cdot \mid X_{it})$, and (ii) she decides whether to switch GP and, if so, to which one. A patient might pay attention due to an event we observe, such as a recent move, or for reasons we do not observe, such as a health event or an interaction with their GP. We model the attention rate p^{λ}_{it} as a function of observables X_{it} and assume attention follows a memoryless Poisson process, so paying attention at time s does not predict attention at $t > s$ conditional on X_{it}.

We model patient choices as forward-looking but myopic. Patients consider the waiting time for each GP, but they do not anticipate future preference changes or switching opportunities. This is consistent with the facts that switch requests are infrequent within patient and that most patients do not fully exploit strategic opportunities even within the relatively simple current waitlist system. If patient i must wait $T \geq 0$ periods to be assigned to GP j, the net present value of requesting this GP is

$$\int_{\tau=0}^{T} e^{-\rho\tau} v_{ij_0} d\tau + \int_{\tau=T}^{\infty} e^{-\rho\tau} v_{ij} d\tau = \frac{1}{\rho}\left[v_{ij_0} + e^{-\rho T}(v_{ij} - v_{ij_0})\right]. \tag{1}$$

Equation 1 shows that the value of switching to another GP can be decomposed into two parts. The first is the value of being forever assigned to their current GP j_0. The second is the *incremental* value of being assigned to GP j instead of j_0 at some point in the future, discounted by waiting time T. Only the second term depends on the chosen GP j.

Let $\mathcal{J}_{it}$ denote patient i's choice set at time t. It remains to specify their information and beliefs about waiting time. The online GP choice interface displays whether each GP has open slots at time t and, if not, the number of individuals on the waitlist for each GP. Let $\mathbf{w}_{it} = \{w_{i1t}, ..., w_{iJt}\} \in \mathbb{N}^J_+$ denote patient i's position if she joined each waitlist. If a GP

[22] In practice, because our data are at the monthly level, we estimate monthly attention probabilities and assume that patients can only be attentive once within a month. Patients do in fact make GP selections continuously throughout the month, and in counterfactual simulations we simulate within-month arrival times as priority tiebreakers and construct the state of the queue at the moment each patient is attentive.

has open slots or i is already on their panel, i's waitlist position is zero.[23] Patient i's choice problem can then be written

$$\max_{j \in \mathcal{J}_{it}} \ \mathbb{E}\left[e^{-\rho T_{ij}} \mid \mathbf{w}_{it}\right] (v_{ijt} - v_{ij_0t}) \,. \tag{2}$$

This choice problem maximizes the value of being assigned to GP j after some waiting time T_{ij}, accounting for uncertainty in waiting time given current waitlist lengths and acknowledging that the patient will remain with her current GP while waiting.

This formulation has several implications. First, there is no explicit cost of waiting. The distaste for waiting time arises only due to exponential discounting.[24] This in turn implies that an attentive consumer will always request to switch to another GP as long as there is *some* other GP in the choice set that delivers higher flow utility than the patient's current GP, regardless of wait time. Our model thus interprets any switch request in the data as implying both that (i) the consumer received an attention shock, and (ii) the requested GP is preferred to the current GP. Any patient who does not request to switch, on the other hand, may be either simply inattentive, or else attentive but prefer the current GP to all others. A second implication of our model is that patients do not necessarily request to switch to their most-preferred GP, in the sense of delivering the highest flow utility. A patient may choose a "second choice" GP with a shorter waitlist in order to wait for less time. Finally, an attentive patient who is relatively satisfied with her current GP is less likely to request to switch. But conditional on requesting to switch, such a patient is more likely to be willing to wait for her most-preferred GP, due to a high "outside option" while waiting.

Discussion of Modeling Choices. The goal of our model of attention and GP choice is to predict how patients might change their behavior when faced with different waitlist lengths and beliefs about waiting time under alternative waitlist mechanisms. A number of our modeling choices warrant specific discussion.

First, we choose to focus on inattention rather than switching costs as the explanation for why GP switch requests are rare (only 19 percent of patients ever request to switch GP during 2017-2019). This choice is not without loss of generality. A model with switching costs would predict that more patients would request to switch should aggregate waiting times fall,

[23] Note that w_{ijt} is i-specific because the patient may already be sitting on a waitlist and would retain their position on that list.

[24] See, e.g. Agarwal et al. (2021), Waldinger (2021), and Verdier and Reeling (2021).

while a model of exogenous inattention would not.[25] We choose to neutralize this channel and focus only on exogenous inattention because GP switch request rates do not appear to respond to local changes in aggregate waiting times. Specifically, we test in the data whether the number of switch requests responds to short-run changes in the average waitlist lengths of all nearby GPs, exploiting a technical change Norway made to its reassignment algorithm during our sample period. Appendix B.5 presents this analysis. We do not find any evidence that patients are more likely to request to switch when waitlists for GPs in their region are particularly short. While we cannot rule out the possibility of a longer-term increase in switch requests as patients become aware that the mechanism has improved, the evidence points toward focusing on modeling *which* GP a patient requests, rather than the decision to switch at all.[26]

Second, our formulation of the choice problem rules out behaviors that would be optimal if patients were fully forward-looking. In particular, patients do not anticipate future preference changes and switching opportunities; they choose a GP as if it will be permanent. This assumption is relatively innocuous for estimation because switch requests are rare at the individual level. The assumption is stronger, however, for a counterfactual mechanism with TTC. Because patients' expected waiting times will depend on whether their current GP is oversubscribed, they may therefore consider not only a GP's current value, but also its future "trading value." In our view, it is unlikely that patients would systematically engage in this type of behavior. For one thing, a GP's trading value is limited by the fact that any gain from modifying one's chosen GP would likely be realized far into the future. Further, patients often fail to fully exploit the dynamics even within Norway's relatively simple current system. For example, it might be optimal for a patient to simultaneously switch to an open GP *and* join the waitlist for an even more preferred GP, which is allowed under the current mechanism. While there are instances of this type of behavior in the data, it is rare—simultaneous open-switches and waitlist joins occur in only 3 percent of patient-months in which a switch request was made.

Finally, we ignore the value of a long-term relationship with your GP. This is one of the motivations for Norway's "patient panel" system, and is surely present to some degree. In

[25] Several papers have considered both exogenous inattention and endogenous switching behavior as explanations for persistent choices (Ho, Hogan and Scott Morton, 2017; Hortacsu, Madanizadeh and Puller, 2017; Abaluck and Adams-Prassl, 2021; Heiss et al., 2021).

[26] Note that in our counterfactual simulations, the number of switch requests is still endogenous because the mechanism may change patients' current GPs when they consider switching. If a patient's current GP is preferable to all others when they are attentive, they will not request to switch.

principle, we could allow a patient's taste for their current GP to depend on the length of the relationship. However, our ability to credibly estimate this object is limited by the fact that we rarely observe the same patient switching multiple times after being assigned to different GPs. Our counterfactual simulations suggest that the rate of switch requests would change very little under alternative mechanisms, even though patients' specific GP choices would adjust in equilibrium. We therefore believe that our proposed mechanism design changes would have limited impact on patient-GP relationship capital.

III.B Waiting Time Beliefs

Waiting time beliefs are a key input to the demand model, operating through the expected discount factor in Equation (2). Modeling beliefs is a key challenge in empirical market design, where market participants often do not have direct access to the information they need to understand the payoffs from different actions. In our setting, patients can easily observe the length of each GP's waitlist, but must infer how this would map into a waiting time. We propse a tractable model of beliefs that approximates the structure of a first-come, first-served (FCFS) queue in a way that depends on a small number of parameters which can be directly estimated from the data. Combined with the choice model, this belief model allows us to translate patients' observed responsiveness to waitlist lengths under the current system to similar responses under alternative mechanisms.

Beginning in waitlist position s, waiting time can be thought of as the sum of the time it takes to move from position s to position $s-1$; from $s-1$ to $s-2$; and so on up to the time from position 1 to being assigned. The time each step takes will depend on the rate at which slots become available on the GP's panel (through departures of incumbent enrollees), and the rate at which patients higher on the waitlist abandon it before being assigned. Incumbent enrollees may depart their panel in the event of death, emigration, or a switch to another GP. Patients standing on waitlists may abandon the waitlist in the event of death, emigration, removing themselves from the waitlist, or switching to a new waitlist.

We assume that patients perceive that slot vacancies and waitlist abandonments follow independent Poisson processes in continuous time. Each slot on GP j's panel becomes vacant at exponential rate η_j, and each patient waiting for GP j abandons the waitlist at exponential rate κ_j. If these processes are independent, then the time t_s it takes to move from position s to $s-1$ follows an exponential distribution with parameter $N_j\eta_j + (s-1)\kappa_j$, where N_j is GP j's panel cap. For a given patient, t_s and $t_{s'}$ are independent for $s \neq s'$. A patient's expected total waiting time when entering GP j's waitlist at position w can then be written as

$$\mathbb{E}[T_j \mid w] = \mathbb{E}\left[\sum_{s=1}^{w} t_s\right] = \sum_{s=1}^{w} \frac{1}{N_j\eta_j + (s-1)\kappa_j},$$

and the expected discount factor as

$$\mathbb{E}\left[e^{-\rho T_j} \mid w\right] = \prod_{s=1}^{w} \frac{N_j\eta_j + (s-1)\kappa_j}{\rho + N_j\eta_j + (s-1)\kappa_j}. \tag{3}$$

Equation (3) embeds two primary simplifications. First, it assumes assignments occur in continuous time at the moment vacancies become available. In practice, assignments are processed at the end of each month, and at least 10 slots must be available before any patients are assigned from the waiting list.[27] Second, it assumes that patients only consider the length of each waitlist in isolation when forming waiting time predictions. In principle, the fact that GP k's waitlist is unusually long—or that all nearby GPs have long waitlists—may predict how quickly GP j's waitlist will move. Given that in practice patients have limited information about the mapping from waitlist lengths to waiting times, we believe our belief model is a reasonable approximation.

IV Estimation

IV.A Sample and Parameterization

Geographic Subsample. For computational tractability, we estimate demand and conduct counterfactual simulations using a geographic and time-period subsample of our data. Geographically, we restrict to patients residing in the Trondelag region. Trondelag is attractive for this purpose because it is a populous region with a major population center (Trondheim), but the region boundaries are sparsely populated.[28] Its population accounts for about 8 percent of the country. In terms of time period, we focus our primary analyses on the 12 month period from December 2018 to November 2019, during which GP waitlists were well-established and active. Appendix Table A.2 provides a comparison of key demographic characteristics

[27] This is done to provide a buffer in case there are multiple births to patients already on the GP's panel.

[28] Only 1.7 percent of residents of Trondelag enroll with a GP outside of the region (compare to 7.7 percent of residents of the Oslo region). Of the 11 administrative regions (*fylke*) in Norway, Trondelag has the lowest outside-region GP enrollment. Appendix Figure A.4 provides a map of Norway and of Trondelag.

between Trondelag and all of Norway over this period. Nothing about the two areas appears substantially different, with the exception that Trondelag is on average more rural.

In addition to geographic and time period restrictions, we also make a number of restrictions relating to the definition of patient choice sets. Our baseline choice set definition is a driving time radius of 60 minutes around a patient's municipality of residence.[29] Details about the construction of the demand estimation sample are provided in Appendix A.2. Our final demand estimation sample represents 4,213,049 patient-months (379,330 unique patients) and 457 unique GPs.

Parameterization. We parameterize the attention and GP choice models as follows.

$$\lambda_{it} \overset{iid}{\sim} \text{Bernoulli}(p^{\lambda}(X_{it})) \tag{Attention}$$

$$v_{ijt} = -d_{ijt} + \delta_j + X_{ijt}\beta + \epsilon_{ijt} \;\;|\;\; \lambda_{it} = 1 \tag{GP choice}$$

$$\epsilon_{ijt} \overset{iid}{\sim} N\left(0, \sigma_{\epsilon}^2(X_{ijt})\right)$$

$$\rho_{it} = \rho \tag{Discount factor}$$

The attention shock λ_{it} is drawn iid each period according to a probability $p^{\lambda}(X_{it})$ that depends on patient observables. Motivated by reduced form evidence in Appendix B.4, we include patient demographics (age, gender, permanent residency status), whether the patient has recently or will imminently move, and if so, the distance of the move. An attentive patient ($\lambda_{it} = 1$) draws new taste shocks for each GP and makes a GP choice. Inattentive patients retain their taste shocks from the previous period. The flow payoff v_{ijt} from being assigned GP j depends on the travel time d_{ijt} between patient i's municipality of residence and the GP's office; a GP fixed effect δ_j capturing the common component of j's desirability, including any unobserved factors (Berry, Levinsohn and Pakes, 2004); interactions $X_{ijt}\beta$ between patient and GP characteristics, capturing common components of patient-GP specific match value; and the idiosyncratic taste shock. Again motivated by the reduced-form evidence, X_{ijt} includes indicators for whether the patient and GP are of the same gender and same age, and we allow the value of age/gender homophily to vary across patient demographics. We allow the variance of the idiosyncratic taste shock to vary by patient age and residency status, in effect allowing variation in the distaste for travel time along these dimensions.

Random utility models require a location and a scale normalization. Our scale normalization

[29]This definition is motivated by the fact that 97 percent of patients enroll with a GP within 60 minutes. Our estimates are not sensitive to adjusting the choice set definition between 45 and 90 minutes.

is to set the distaste for travel time to 1. Travel time thus acts as a numeraire in the absence of prices, as is common in empirical market design applications (Abdulkadiroğlu, Agarwal and Pathak, 2017; Agarwal and Somaini, 2018). Our location normalization is to set the fixed effect for one GP to zero. This leaves us with the following model parameters to estimate: $\{\rho, \delta, \beta, \sigma_\epsilon, p^\lambda\}$.

IV.B Estimation Procedure & Identification

We estimate the model in two steps. We first estimate the parameters governing the formation of patient beliefs using the empirical analogs of the objects described in section III.B. Using this model of patient beliefs, we then jointly recover the attention and preference parameters that best describe the observed data given our model.

Waiting Time Beliefs. Consumers have concrete information about waitlist lengths and panel sizes, but have limited information about other factors. We therefore assume that $(\eta_j, \kappa_j) = (\eta, \kappa)$ and estimate the latter on a per-month basis across all waiting list-month observations that are in the choice set of our demand estimation sample. Out of 5,140 total GP panel-months considered (representing 457 unique GPs), 2,161 were panel-months in which the panel had a waitlist (representing 298 unique GPs). Among these panel-months, we then calculate the average panel slot vacancy rate and the average waitlist abandonment rate, yielding belief parameters:

$$\eta = 0.0018 \qquad \text{(Panel slot vacancy rate)}$$

$$\kappa = 0.0170 \qquad \text{(Waitlist abandonment rate)}$$

The average waiting list length over these months was 28, and the average panel cap was 1,084. Using our formula for expected waiting time, this panel cap and waitlist length coupled with our belief parameters would imply an expected waiting time of 12.8 months. For comparison, the average elapsed wait time among patients on waitlists as of December 2019 was 7.9 months (but note that waitlist lengths were still growing rapidly at this point, such that this does not reflect the full realized waiting time that those patients ultimately experienced).

Attention and Preferences. We estimate the attention and GP choice model using Markov Chain Monte Carlo (MCMC) methods. We use a Gibbs' sampler with data augmentation to draw attention shocks λ_{it} and flow utilities v_{ijt} from their posterior distributions, and a

Metropolis-Hastings step to update the discount factor ρ (McCulloch and Rossi, 1994; Gelman et al., 2013). We assume conjugate priors of $(\delta, \beta) \sim N(\mu_0, \Sigma_0)$, $p^\lambda \sim Beta(\alpha, \varphi)$, and $\sigma_\epsilon^2 \sim IW(\Psi_\epsilon^0, \nu_\epsilon^0)$. The steps of the Gibbs' sampler can be written as follows for iteration b, where each step also conditions on patients' observable characteristics (Z) and choices (y):

(i) $\delta_b, \beta_b \mid \mathbf{v}_{b-1}, \sigma^2_{\epsilon,b-1}, \mu_0, \Sigma_0$

(ii) $\sigma^2_{\epsilon,b} \mid \mathbf{v}_{b-1}, \delta_b, \beta_b, \Psi_\epsilon^0, \nu_\epsilon^0$

(iii) $\rho_b \mid \delta_b, \beta_b, \sigma^2_{\epsilon,b}, y, Z$

(iv) $\lambda_b \mid p^\lambda_{b-1}, \delta_b, \beta_b, \sigma_{\epsilon,b}, y, Z$

(v) $p^\lambda_b \mid \lambda_b, \alpha, \varphi$

(vi) $\mathbf{v}_b \mid \lambda_b, \delta_b, \beta_b, \sigma^2_{\epsilon,b}, \rho_b, y, Z$

Appendix C provides additional details on the updating steps. Though our estimator is Bayesian, it is asymptotically equivalent to maximum likelihood estimation (see, e.g., Van der Vaart 2000, Theorem 10.1, Bernstein von Mises) and computationally less demanding.[30] We interpret the posterior means and standard deviations in a frequentist manner for the purposes of inference.

Identification. We can think of identification in three steps: identifying the distribution of (i) flow payoffs, (ii) the discount factor, and (iii) the attention parameters. Beliefs are assumed known.

First, suppose the discount factor ρ is known and attention is observed. The distribution of flow payoffs is non-parametrically identified by variation in travel time between patients' residences and GP offices. This argument treats travel time as a special regressor (Berry and Haile, 2014; Agarwal and Somaini, 2018) and requires that unobserved determinants of preferences for GPs are uncorrelated with patients' proximity to GP offices, conditional on observables. The key economic assumptions are that patients do not choose where to live based on access to primary healthcare, and that GPs do not locate their offices close to where patients live who particularly value seeing those GPs. We believe this is plausible in our context.

Second, given the distribution of flow payoffs, the discount factor is identified by the sensitiv-

[30] Because discounting is multiplicative, the full likelihood would not have a tractable closed-form even if we assumed ϵ was distributed Type-1 Extreme Value.

ity of patients' choices to waitlist lengths (Waldinger, 2021).[31] A key identification challenge is that more desirable GPs will tend to have longer waitlists. However, because panel vacancies and waitlist departures are stochastic events, queue lengths naturally fluctuate around their long-run averages due to events outside the control of a patient considering switching. This generates exogenous variation in the relative waitlist lengths of different GPs for patients who consider switching GPs at different times. A key identifying assumption is that patients do not time *when* they request to switch based on the waitlist lengths of specific GPs. In a FCFS queue, when there is no direct cost of standing on a waitlist, there is no benefit to delaying a switch request until a patient's desired GP's queue is unusually short. We therefore view this threat as unlikely. It is also possible that a component of GP desirability is time-varying, so that a GP's waitlist tends to grow when the GP becomes more desirable. To the extent that this occurs, this would lead us to underestimate the discount factor.

Finally, the attention parameters are separately identified by attention shifters that are uncorrelated with preferences for specific GPs. In the ideal experiment, imagine a group of patients who pay attention with probability one (Abaluck and Adams-Prassl, 2021). If their preferences are drawn from the same distribution as the general population, their choices identify the other model parameters. We can then recover attention probabilities for the remaining patients by comparing their frequencies of switch requests and GP choices to those of always-attentive patients. In the data, we observe that patients who move a long distance are much more likely to request to switch GPs than patients who move shorter distances or stay put. Our identifying assumption is that conditional on other observables, the distribution of GP preferences is uncorrelated with when or how far patients move. In practice, many patients wait years to switch GPs even after a long move, so we rely on parametric assumptions to jointly estimate patients' preferences and attention probabilities.

IV.C Estimates

We estimate three specifications. All specifications allow attention to differ flexibly by patient age, gender, and whether the patient is about to or recently moved. Flow payoffs are parameterized similarly to the conditional logistic regressions in Section II.B, with interactions between patient and GP age and gender in all specifications. The second specification adds

[31] Though we estimate a common discount rate for all patients, in principle, one could allow it to depend on any observed characteristics. In practice, it was difficult to obtain precise estimates of ρ for different subgroups, and we could not reject that our point estimates were the same among subgroups in the specifications we tried.

GP fixed effects, and the third specification allows the standard deviation of the idiosyncratic taste shock to vary by patient age and permanent residency status. All specifications estimate a common monthly discount rate ρ, reported as an annual rate.

Table 3 presents the parameter estimates. Column (1) estimates considerable sensitivity to waiting time, with an annual discount factor slightly below 0.95. When GP fixed effects are added in columns (2) and (3), this value falls to between 0.90 and 0.91, reflecting the fact that more desirable GPs have longer waitlists. As in the conditional logits, we estimate considerable homophily by gender and age. Based on estimates in column (1), a female patient under age 45 would travel 7.3 minutes farther than a male patient under 45 to see a female GP (6.3 minutes for a female/male patient over 45). Patients under 45 would travel about one minute farther than a patients over 45 to see a GP under 45. Adding GP fixed effects and heterogeneity in the variance of the taste shock to the model leaves the estimates of age and gender homophily nearly unchanged.

Column (3) allows the variance of the idiosyncratic taste shock to vary by patient age and residency status, which can be equivalently thought of as allowing for heterogeneity in the distaste for travel time. We find that the distaste for travel time is considerably higher for older permanent residents than for other patients. In both columns (2) and (3), we find considerable variation in overall GP desirability as well as in the value of the idiosyncratic taste shock. Column (2) estimates a standard deviation of GP fixed effects of 31 minutes, and a standard deviation of idiosyncratic shock of 12.6 minutes. In column (3), the standard deviation of GP fixed effects falls to between 16 and 24 minutes, and a standard deviation of idiosyncratic shock to between 9 and 12 minutes.[32]

Table 4 presents parameter estimates for the attention model. Estimates are very similar across specifications, so we focus on column (3). Our estimates imply that non-movers consider switching GPs far less often than patients who have recently moved, consistent with observed switching patterns. Among non-movers, temporary residents pay attention most often——1.084 percent chance per month (approximately once every 7.5 years)——while older men consider switching just once every 25 years.

Patients who have moved consider switching an order of magnitude more often. We estimate

[32]Variance in the idiosyncratic shock may in part reflect the fact that we observe only a patient's municipality of residence rather than their exact address, introducing measurement error in travel time. Using data from an earlier time period in which we had exact address, we investigated whether measurement error in travel time from observing municipality rather than exact address is correlated with other patient characteristics, and found essentially no relationship.

Table 3. Preference Parameter Estimates

Variable	(1) β	(1) SE	(2) β	(2) SE	(3) β	(3) SE
Annual Discount Factor	0.944	0.002	0.904	0.007	0.908	0.007
Travel time (minutes)†	−1.000		−1.000		−1.000	
S.D. GP Fixed Effects			31.008	4.962		
S.D. GP Fixed Effects, Temp. res.					23.092	1.776
S.D. GP Fixed Effects, Perm. res. age $\leq$ 45					24.195	1.722
S.D. GP Fixed Effects, Perm. res. age > 45					16.400	1.264
Male GP	−3.036	0.365				
× Perm. res. female, age 16–45	−1.300	0.414	−2.448	1.051	−2.886	0.425
× Perm. res. female, age 45+	−0.608	0.424	−0.922	1.014	−1.170	0.477
× Perm. res. male, age 16–45	5.986	0.421	4.967	0.997	4.054	0.474
× Perm. res. male, age 45+	5.754	0.477	5.947	1.018	2.761	0.686
GP age 45+	−2.184	0.404				
× Perm. res. female, age 16–45	−0.446	0.441	−0.821	1.090	−1.787	0.394
× Perm. res. female, age 45+	0.499	0.440	0.330	1.053	−0.583	0.482
× Perm. res. male, age 16–45	−0.404	0.460	−0.309	1.071	−1.496	0.424
× Perm. res. male, age 45+	0.267	0.558	0.937	1.069	−0.713	0.632
S.D. idiosyncratic shock	14.885	0.161	12.643	0.334		
S.D. idiosyncratic shock, Temp. res.					12.296	0.314
S.D. idiosyncratic shock, Perm. res. age $\leq$ 45					12.894	0.470
S.D. idiosyncratic shock, Perm. res. age > 45					8.738	0.378

Notes: The table reports parameter estimates from the GP choice model described in Section IV. We simulate 40,000 draws from the Markov chain and drop the first 20,000 for each specification. The table reports the mean and standard deviation of the remaining draws as the point estimate and standard error of each parameter (respectively). Columns (2) and (3) include a fixed effect for each GP, with one normalized to zero. Column (3) allows the standard deviation of the idiosyncratic shock to differ by residency status and age. †By normalization

separate attention probabilities for short and long moves, where short means less than 30 minutes' drive time between original and destination municipality, and long means over 30 minutes. We also allow attention to differ by whether the move is imminent (occurring this month or next month) or recent (in the past 6 months), reflecting the fact that we see a large immediate impact of moving on the probability of switching GPs in the data (c.f. Appendix Table B.2). Finally, we allow mover attention to depend on gender and residency status. Movers exhibit stark differences in attention probabilities along all three dimensions. Patients are most attentive in the month prior to or of a move. A temporary resident moving over 30 minutes this or next month has an 18.59 percent chance of considering switching GPs. The probability remains high, but falls to 7 percent per month in the six months following the move. Permanent residents exhibit a similar pattern, but with lower attention probabilities.

Table 4. Monthly Attention Probability Estimates

Variable	(1) p^λ	(1) SE	(2) p^λ	(2) SE	(3) p^λ	(3) SE
No Recent or Imminent Move						
× Temporary resident	1.21	0.02	1.33	0.02	1.08	0.02
× Perm. res. female, age ≤ 45	0.75	0.01	0.85	0.01	0.82	0.02
× Perm. res. female, age > 45	0.47	0.01	0.47	0.01	0.47	0.01
× Perm. res. male, age ≤ 45	0.44	0.01	0.44	0.01	0.48	0.03
× Perm. res. male, age > 45	0.28	0.01	0.28	0.01	0.33	0.01
Moved ≤30 minutes, this or next month						
× Temporary resident	6.92	1.13	6.86	1.13	6.93	1.13
× Perm. res. female	4.63	0.38	4.63	0.37	4.63	0.39
× Perm. res. male	3.35	0.33	3.35	0.33	3.35	0.33
Moved ≤30 minutes, prev. 6 months						
× Temporary resident	3.43	0.55	3.42	0.55	3.41	0.57
× Perm. res. female	2.44	0.18	2.45	0.18	2.45	0.17
× Perm. res. male	5.99	0.27	1.71	0.15	1.71	0.15
Moved >30 minutes, this or next month						
× Temporary resident	18.50	2.27	18.51	2.26	18.59	2.32
× Perm. res. female	8.85	0.67	8.84	0.66	8.83	0.65
× Perm. res. male	6.86	0.59	6.82	0.58	6.84	0.61
Moved >30 minutes, prev. 6 months						
× Temporary resident	7.28	1.13	7.33	1.17	7.28	1.12
× Perm. res. female	3.83	0.28	3.82	0.28	3.83	0.29
× Perm. res. male	2.86	0.24	2.85	0.24	2.85	0.24

Notes: The table reports parameter estimates from the attention model. Parameters represent the probability a patient is attentive in a given month, reported in percentage points. These parameters are estimated jointly with the preference parameters reported in Table 3. The categories are mutually exclusive and exhaustive. A patient has *No Recent or Imminent Move* if they did not move across municipalities in the six months prior or one month after the current month. After a patient has been attentive during a given move spell, they are treated as a non-mover during the rest of that move spell.

Distance of move is also highly predictive. A female permanent resident moving less than 30 minutes has only a 22 percent cumulative attention probability in the 8 months around that move. If the move were over 30 minutes, this rises to 34 percent. Attention around move events will be important for counterfactual simulations, because patients are most attentive when they are also most mismatched with their current GP.

V Counterfactual Simulations and Welfare

We now use our estimates to predict equilibrium assignments under alternative waitlists mechanisms. Section V.A describes the simulated dynamic economy. Section V.B then formally defines our counterfactual mechanisms of interest, Section V.C defines an equilibrium, and Section V.D presents the results.

V.A Simulated Dynamic Economy

We consider an economy with a finite set of patients (agents) and GPs (objects). The set of patients is given by $I = \{i_1, ..., i_n\}$. Patients have types $x \in \mathcal{X}$. The set of GPs is given by $J = \{j_1, ..., j_m\}$. GPs have types $z \in \mathcal{Z}$. We treat $\mathcal{X}$ and $\mathcal{Z}$ as finite sets. An allocation $\mu : I \to J$ is a many-to-one mapping of patients to GPs, such that each patient is assigned to exactly one GP. The allocation prevailing at a given time t is denoted by μ_t.

GP types include age, gender, office location, and panel cap size. All GP characteristics are fixed over time. Patient types include demographics and location of residence. There are five possible demographic types: {Perm. res. female ≤ 45, Perm. res. female > 45, Perm. res. male ≤ 45, Perm. res. male > 45, Temp. res.}, and 45 possible locations of residence (municipalities in the Trondelag region). The combination of patient municipality and GP office location pin down a travel time between every patient type and every GP.

Patient characteristics evolve over time according to a stationary Markov process $M : \mathcal{X} \to \mathcal{X}$. Patients that transition from old to young experience a "rebirth" procedure in which they are removed from their current GP and reassigned to the current GP of a randomly selected young woman living in their same location of residence. The simulation therefore allows for panel vacancies to arise naturally, but given the standard practice of enrolling babies with their mother's GP, less frequently on GP panels with many young women.

We initialize the set of patients and GPs in the economy based on the distribution of patient and GP types observed Trondelag region in December 2019. There are 371,536 patients and 425 GPs. We calibrate the patient type transition process M based on type transitions observed in Trondelag over the period 2017–2019. Appendix Table D.1 provides summary statistics on the set of patients and GPs in the simulated economy. Appendix D.1 provides additional details on the construction of our initial conditions as well as on the transition process M.

Simulation Procedure. In each period, the following steps occur (in order):

1. (*Demographic transitions*) Patients draw a new type. Patients that die are removed from their current GP as well as any waitlists they are on, and added to the GP panel of the mother to whom they are reborn.

2. (*Attention*) Patients draw attention shocks according to the probabilities $p^\lambda(x)$ estimated in Table 4.

3. (*GP Choice*) Attentive patients sequentially arrive to the mechanism. Upon entry, they consider all GPs, draw new preferences (ϵ_{ij}), formulate flow indirect utility for each GP (v_{ij}) (based on parameter estimates in Table 3), and report their GP request(s) to the mechanism.

4. (*Matching Algorithm*) The patient-GP matching algorithm is executed. Patients that are successfully reassigned exit the mechanism. All other patients in the mechanism remain there until the following period.

Realizations of the demographic transition process, the receipt of attention shocks, and draws of idiosyncratic preferences are held fixed across simulations of alternative mechanisms. Appendix Table A.3 summarizes these realizations.

V.B Allocation Mechanisms

Let μ_0 be the allocation at the start of a given period. Each GP j has capacity N_j, a set of currently enrolled patients $\mu_0^{-1}(j)$, and a set of patients w_j on their waitlist. Each patient on a waitlist has waited a length of time t_i. Let $\succ_j$ denote GP j's (strict) preferences over patients, which encode the priority rules of the mechanism. Under a **first-come, first-served (FCFS)** priority rule, patients that have waited longer have higher priority: $\forall\, l, k \notin \mu_0^{-1}(j)$, $l \succ_j k$ iff $t_l > t_k$. Under a priority rule that **respects endowments**, incumbent patients always have higher priority than non-incumbent patients at their current GP: $\forall\, l, k \in I$, $k \succ_j l$ if $k \in \mu_0^{-1}(j)$ and $l \notin \mu_0^{-1}(j)$. All priority rules we consider will respect endowments.

While GPs (objects) are not strategic, patients (agents) may be. An attentive patient submits a rank-order preference list (ROL) R_i. In all of our mechanisms of interest, patients may join at most one waitlist, so R_i has length at most two: $(j, \mu_0(i))$. The requested (waitlist) GP is ranked first, and the current GP is ranked second. A *matching algorithm* ϕ maps a set of patient-reported ROLs $\mathbf{R}$, GP priorities $\succ$, and panel caps $\mathbf{N}$ into an allocation μ. An *allocation mechanism* is defined as the triplet $[\mathbf{N}, \succ, \phi]$, in combination with rules regarding the maximum length of patients' ROLs and how often the matching algorithm is run. Our counterfactuals change both the priority rule $\succ$ as well as the matching algorithm ϕ applied each period. Patients' reported ROLs respond endogenously to these changes.

While the matching algorithm is run only periodically (once per month), an allocation mechanism operates in continuous time. Attentive patients arrive continuously and submit ROLs. They then remain in the mechanism until they can be successfully reassigned (or they die). If they receive a subsequent attention shock while still in the mechanism, they have the option to change their ROL (switch waitlists), resetting their wait time priority. So long as GP priorities respect endowments and patients can include their current GP at the end of their ROL, it is ex-post individually rational for attentive patients to participate in the mechanism. Further, under FCFS priorities—or any other priority rule that is increasing in waiting time—a patient cannot benefit by delaying when they request to switch. Attentive patients therefore should not be strategic in terms of *whether* to participate in the mechanism, only in their reported preferences.

We consider three primary counterfactual allocation mechanisms, in addition to the status quo mechanism. First, we run the TTC algorithm on top of Norway's existing waitlists, just as in our original simulations from Section B.1, but where we can now endogenize patient responses. We then consider two alternatives, intended to address the distributional and fairness concerns that—as we saw in the naive simulations—may arise under TTC. Each of these mechanisms, as well as Norway's status quo mechanism, are described in detail below.

Waitlists. This is the allocation mechanism currently used in Norway. Priorities are FCFS. The matching algorithm used is the following:

> *Step 1:* For each GP j, let $O_j = N_j - |\mu_0^{-1}(j)|$ be the number of open slots on j's panel. Reassign the O_j highest-priority patients on j's waitlist to j's panel; remove each of these patients from their current GP's panel and from the waitlist.
>
> *Step k:* Repeat Step 1 for the panels and waitlists resulting from Step $k-1$.

The algorithm terminates if no patients are reassigned in a step. When this occurs, there are no patients waiting for GPs with open slots.

Waitlists with Top-Trading Cycles (TTC). This is our main proposal for Norway's GP allocation system. Priorities are still FCFS. The matching algorithm used would be to first run the Waitlists algorithm, and then run the TTC algorithm. The TTC algorithm works as follows. Each GP begins with pseudo-capacity $\tilde{N}_j$ equal to their number of open panel slots plus the number of their *current* patients who are currently in the mechanism hoping to switch away to another GP.

> *Step 1:* Each patient "points to" their preferred GP according to R_i, and each GP points

to their preferred patient according to $\succ_j$. There is at least one cycle, i.e., an ordered list $\{i_1, j_1, i_2, j_2, ..., i_k, j_k\}$ where i_1 points to j_1, j_1, points to i_2, ... , and j_k points to i_1. Further, each patient and GP can be part of at most one cycle. Each patient in a cycle is reassigned to the GP they point to and is removed from the algorithm. Each GP in a cycle has their pseudo-capacity reduced by one, and is removed from the algorithm when their pseudo-capacity falls to zero.

Step k: Repeat Step 1 with the *remaining* patients and updated GP pseudo-capacities from the end of Step $k-1$.

The algorithm terminates when no patients remain in the algorithm.

Waitlists with Top-Trading Cycles and Priority (TTCP). This mechanism is identical to TTC, but modifies priorities $\succ$ so that patients with undersubscribed GPs are prioritized above patients with oversubscribed GPs (while still respecting endowments). Formally, $\forall\, l, k \notin \mu_0^{-1}(j), l \succ_j k$ if l currently has an undersubscribed GP and k currently has an oversubscribed GP. Among patients with the same over/undersubscribed status, there is FCFS priority. We classify a patient's current GP as over/undersubscribed at the moment they enter the mechanism.[33] A GP is undersubscribed if it has at least one open slot on its panel.

TTCP attempts to address some of the adverse distributional consequences from introducing TTC, while preserving its benefits. Rather than preventing waiting patients from exchanging their GPs, TTCP prioritizes patients who are unlikely to benefit from cycles for other vacancies.

Waitlists with Deferred Acceptance (DA). This mechanism is identical to Waitlists, but replaces the matching algorithm used with the patient-proposing deferred acceptance (DA) algorithm. Again, let $\tilde{N}_j$ denote GP j's pseudo-capacity. The DA algorithm proceeds as follows:

Step 1: Each patient "proposes to" their preferred GP according to R_i. Each GP j provisionally accepts proposals from its $\tilde{N}_j$ most preferred patients according to $\succ_j$, and rejects any remaining proposals.

Step k: Any patients who were rejected in the previous round propose to their most-preferred GP (according to R_i) who has not yet rejected them. Each GP j provisionally

[33] While it is possible that a GP could transition to/from being over/undersubscribed while their patient remains in the mechanism, these transitions rarely occurred in our simulations, and capturing patient beliefs about this possibility would be overly complex.

accepts proposals from its $\tilde{N}_j$ most preferred patients according to $\succ_j$, and rejects any remaining proposals.

The algorithm terminates when no patient is rejected.

Relative to Waitlists, the key advantage of DA is to allow trades among patients at the front of their respective waitlists. Relative to TTC, however, an attractive property of DA is that, like Waitlists, it strictly respects FCFS waiting time priority. No patients can "jump" to the front of the queue and be reassigned to a GP for whom another patient was waiting longer. This idea reflects DA's well-known property of *stability* (Gale and Shapley, 1962), which is also known as *elimination of justified envy* (Abdulkadiroğlu and Sönmez, 2003). Further, because patient-proposing DA yields the patient-optimal stable match with respect to reported preferences, it represents the best *any* algorithm in this class can do without violating FCFS priority (i.e., generating envy). Nevertheless, by strictly respecting waiting time priority, DA may substantially limit trading opportunities relative to TTC.

V.C Beliefs, Decisions, and Equilibrium

In all of our counterfactuals, patients solve the choice problem in Equation (2):

$$\max_{j\in\mathcal{J}_{it}} \ \mathbb{E}\left[e^{-\rho T_{ij}} \mid \mathbf{w}_{it}\right](v_{ijt} - v_{ij_0t}).$$

Counterfactual mechanisms will not only change the number of patients on the waitlist at a given time $\mathbf{w}_{it}$, but also patients' beliefs about the speed with which waitlists move, i.e., the expected discount factor function $\mathbb{E}\left[e^{-\rho T_{ij}} \mid \mathbf{w}_{it}\right]$. How these beliefs adjust will determine patients' equilibrium responses to alternative mechanisms. To accommodate the additional complexity of TTC, we must adapt the belief model from Section III.B.

As before, patients have beliefs about how quickly a waitlist moves from the front (η), and how often waiting patients depart the waitlist before reaching the top (κ). Under TTC, beliefs must also account for the fact that a patient can be successfully reassigned before reaching the top of the waitlist if they participate in a cycle. Further, patients ahead in the queue may depart because they are assigned through a cycle, as well as due to exogenous departures.

We modify the belief structure as follows. In each position s, a patient will either (i) move to position $s-1$ due to a waitlist departure or assignment from the front of the queue, or (ii) be reassigned through a cycle. We assume that these two events are perceived to follow independent, memoryless arrival processes. We allow the rate of being assigned through a cycle $\chi_{j_0 s}$ to depend on a patient's position s as well as on whether their current GP j_0 is undersubscribed.[34] Patients are assigned from the top of GP j's waitlist at rate $N_j\eta$. Each patient on the waitlist departs at rate κ, which now includes both abandonments and reassignments through a cycle. A patient's expected discount factor when entering GP j's waitlist at position s can now be written

$$\begin{aligned}\mathbb{E}[e^{-\rho T_{ij}} \mid j_0, s] &= \frac{m_{js} + \chi_{j_0 s}}{\rho + m_{js} + \chi_{j_0 s}} \left(\frac{\chi_{j_0 s}}{m_{js} + \chi_{j_0 s}} + \frac{m_{js}}{m_{js} + \chi_{j_0 s}} \mathbb{E}[e^{-\rho T_{ij}} \mid j_0, s-1] \right) \\ &= \frac{\chi_{j_0 s}}{\rho + m_{js} + \chi_{j_0 s}} + \frac{m_{js}}{\rho + m_{js} + \chi_{j_0 s}} \mathbb{E}[e^{-\rho T_{ij}} \mid j_0, s-1], \end{aligned} \tag{4}$$

where $m_{js} \equiv N_j\eta + (s-1)\kappa$ is the rate at which the patient moves forward in the queue.[35] Equation 4 provides a recursive formula for the expected discount factor at any position s. We parameterize $\chi_{j_0 s}$ such that it equals zero for patients with an undersubscribed GP (who have no chance of participating in a cycle), and is log-linearly related to waitlist position-relative-to-panel cap for patients with an oversubscribed GP: $\chi_{j_0 s} = \mathbb{1}[j_0 \text{ oversub.}] \exp(\chi_0 + \chi_1 \log(s/N))$.[36] Note that if $\chi_{j_0 s}$ is restricted to zero for all patients, this formulation of beliefs collapses to the original beliefs structure relevant for Waitlists and DA.

We compute counterfactual equilibria in which belief parameters are consistent with the waiting times implied by patients' optimal decisions. Appendix D.2 describes our procedure in detail. The algorithm iteratively updates patients' optimal decisions given the state of the simulation algorithm and the belief parameters implied by the simulation, until the belief parameters (and hence decisions) converge. Appendix Table D.2 reports equilibrium beliefs.

[34] In principle, $\chi_{j_0 s}$ could depend in a complex way on the state of the mechanism, including the lengths of all GP waitlists and number of open slots on each GP's panel, as well as the GP the patient has requested. Our simplification strikes a balance between capturing the most important determinants of waiting times and having low complexity.

[35] The intuition behind this formula is as follows. The next event that occurs is either participating in a cycle or moving one position forward in the queue. If these two processes are independent and memoryless, then the next event occurs at exponential rate $m_{js} + \chi_{j_0 s}$. The ratio $\frac{m_{js}+\chi_{j_0 s}}{\rho+m_{js}+\chi_{j_0 s}}$ is the expected discount factor for the time of that event. Given that an event occurs, it is participating in a cycle with probability $\frac{\chi_{j_0 s}}{\chi_{j_0 s}+m_{js}}$, in which case assignment is immediate and the subsequent discount factor is 1. The probability the event is moving up a position is $\frac{m_{js}}{\chi_{j_0 s}+m_{js}}$, and the subsequent discount is simply $\mathbb{E}[e^{-\rho T_{ij}} \mid j_0, s-1]$.

[36] In the TTCP mechanism, we allow patients with undersubscribed GPs to have different beliefs about κ, since any patients ahead of them on a waitlist will also have undersubscribed GPs and thus have no chance of departing the waitlist by participating in a cycle.

V.D Results

Table 5 reports equilibrium outcomes under our primary mechanisms of interest. It also reports outcomes under a benchmark simulation, **No Caps**, in which GP panel caps (and thus all scarcity in the economy) are removed. This benchmark provides an upper bound on the welfare that can be achieved by any mechanism within our framework.[37] We report outcomes from a 5-year window at the end of out simulation period, when the economy has reached a stationary equilibrium.

Table 5. Outcomes under Alternative Mechanisms

	Waitlists	TTC	DA	TTCP	No Caps
GP waitlists					
Pct. of population on a waitlist	9.4	8.9	9.3	9.6	–
Pct. of GPs with a waitlist	82.2	78.3	82.1	79.3	–
Mean E(waittime) | curr. GP undersub.	16.7	22.8	16.7	18.9	–
| curr. GP oversub.	16.7	10.7	16.7	12.2	–
Attentive patient choices					
Mean E(waittime) at chosen GP	16.8	14.1	16.8	15.2	–
Pct. waitlist joins	85.2	84.6	85.2	85.1	–
| curr. GP undersub.	79.0	74.8	79.1	77.2	–
| curr. GP oversub.	86.2	86.3	86.1	86.3	–
True pref. rank of chosen GP	1.79	1.63	1.78	1.63	1.00
| curr. GP undersub.	1.94	2.29	1.93	2.04	1.00
| curr. GP oversub.	1.76	1.52	1.76	1.56	1.00
Realized assignments					
Travel time to current GP, mean (med.)	17.3 (6.5)	16.9 (6.4)	17.3 (6.5)	17.1 (6.4)	16.8 (6.4)
Pct. with same gender GP | young female	59.3	60.3	59.3	60.3	68.7
| young male	61.8	61.2	61.9	61.1	52.6
Welfare					
Flow payoff from current GP, mean (med.)	–†	0.75 (0)	0.01 (0)	0.46 (0)	5.34 (0)
Perpituity equiv. of GP choice, mean (med.)	–†	1.25 (0.8)	0.01 (0.0)	1.08 (0.5)	6.15 (4.2)

Notes: The table reports statistics on outcomes generated in months 392–451 of the simulation, out of 500 total months. Statistics are first computed within month and then averaged across simulation months. E(waittime) is the expected waiting time implied by patient's equilibrium beliefs and current waitlist lengths. True pref. rank of requested GP is the rank of a patient's requested GP in their true flow payoff ordering. †By normalization.

In the long-run stationary equilibrium of Norway's current mechanism (Waitlists), 9.4 percent of patients are on a waitlist, and 82.2 percent of GPs have a waitlist. Patients' expected waiting time to switch to the average GP (including zeros for those without waitlists) would be 16.7 months. Each month, an average of 2,299 patients receive attention shocks and consider switching GPs (c.f. Appendix Table A.3). Among these patients, 85.2 percent choose to join

[37] Of course, allowing GP panel sizes to grow arbitrarily would in reality likely reduce the value of highly-demanded GPs, for example by increasing waiting time for appointment.

a waitlist, while the rest choose a GP with open slots or to remain with their current GP. The average attentive patient expects to successfully obtain their chosen GP after 16.8 months.[38] Despite considerable sensitivity to waiting time, many patients are willing to wait for their first choice GP. The average rank of a patient's requested GP in their true preference list is 1.79.

To quantify the gains from introducing TTC, we measure welfare in two different ways. First, we calculate the flow payoff experienced by every patient in the economy each period. This measure would be relevant to the utilitarian policymaker interested in maximizing the present discounted value of all future payoffs in the economy. It also has the advantage that it relies only on realized assignments. However, in a dynamic economy, a patient's current GP at any given time may differ across mechanisms for two reasons: (i) due to GP switching events (attention spells) that occurred earlier in time, and (ii) due to contemporaneous differences in the ease with which an attentive patient can switch GPs due to changes in equilibrium waiting times. To highlight the latter channel, we introduce an alternative welfare measure that isolates the value of switching opportunities holding a patient's current GP fixed. Specifically, we measure the net present value (NPV) of an attentive patient's choice problem, holding the patient's current GP fixed at its realized value under Waitlists.[39] We then multiply by patients' discount factor ρ (estimated to be 0.0081), yielding the perceived perpetuity flow payoff associated with attentive patients' optimal choice under each mechanism. This measure highlights the gains for patients precisely when they consider switching GPs. It also allows us to compare patient welfare as a function of characteristics that are endogenous to the mechanism, e.g., whether the patient has an over- or undersubscribed GP. For both measures, we normalize welfare under Waitlists to zero for all patients.

Relative to the status quo, introducing TTC reduces waiting times and increases patient welfare. The average attentive patient now successfully obtains their chosen GP after 14.1 months, and a smaller share of patients are on a waitlist at a given time. Measured by mean flow payoffs from realized assignments, patient welfare increases by the equivalent of 0.75 minutes' driving time. This improvement is significant in magnitude (more than 13 percent of the gains under the highly infeasible benchmark of No Caps). More than half is attributable directly to patients being matched with closer GPs (0.4 minutes), with the remainder driven by improved match quality on both observable and unobservable dimensions. The equivalent

[38] Relative to our data, these much longer equilibrium waiting times reflect the fact that the waitlists were still growing rapidly during our sample period.

[39] Appendix D.3 provides additional details.

perpetuity perceived by the average attentive patient is 1.25.[40] These average improvements are reflected in patients' behavioral responses; they are more likely to request their first-choice GP, with the mean flow payoff rank of their chosen GP falling to 1.63.

Distributional Implications. Which patients benefit from TTC, and why? Table 6 summarizes welfare outcomes by patient subgroup (using the NPV measure), comparing the welfare gains under each alternative mechanism. In terms of demographics, the welfare gains from TTC are concentrated among younger patients, especially women (mean increases of 1.5 minutes for young women, 1.4 for young men). Patients who moved in the last 12 months especially benefit (2.3 minutes). This reflects the fact that under TTC, patients are more likely to be able to receive a desirable GP quickly when they are highly mismatched with their current GP. However, patients who have never moved still also benefit (1.0 minutes), reflecting the fact that patient moves are not the only dynamic driving patient-GP mismatch. Patients in both urban and rural areas benefit, but the gains are largest among rural patients (2.1 minutes), who tend to face the longest travel times and thus have the greatest potential for geographic mismatch. Finally, it is worth noting that for most groups, changes in median welfare are also positive and economically significant, suggesting the gains are widely spread rather than concentrated among a minority of patients.

Although TTC benefits many patients, it has particularly uneven distributional consequences along an important dimension—whether a patient's current GP is oversubscribed. Table 5 shows that while waiting times for the average GP decrease by 6 months (from 16.7 to 10.7) for patients with oversubscribed GPs, they *increase* by 6 months (from 16.7 to 22.8) for patients whose GPs are undersubscribed. Patients' GP choices then reflect the disparity along this dimension. Patients with undersubscribed GPs become less selective under TTC, choosing GPs with an average preference rank of 2.29 instead of 1.94 under Waitlists, and joining waitlists less frequently (74.8 percent instead of 79.0 percent). Patients with oversubscribed GPs, in contrast, become more selective, choosing better-matched GPs and becoming more likely to join a waitlist.

These differences in waiting times and behavioral responses are reflected by our NPV welfare

[40] The difference between these numbers reflects two things. First, the NPV measure re-weights patient months relative to the flow payoff measure because it is calculated in *attentive* patient months only. Since attentive patients are weakly improving their situation (relative to being inattentive), welfare gains tend to be larger according to the NPV measure than according to the flow payoff measure. Second, patients are over-optimistic about the flow of welfare gains they will receive over an infinite horizon because in reality they experience subsequent shocks (e.g., moving or dying), the probability of which they do not take into account at the moment of attention.

Table 6. Distribution of Welfare Gains

	Frac. of attn. pats.	TTC	DA	TTCP	No Caps
All attentive patient-months	1.00	1.3 (0.8)	0.0 (0.0)	1.1 (0.5)	6.1 (4.2)
Current GP oversubscribed?					
No	0.13	-0.8 (-0.4)	0.0 (0)	0.0 (0)	6.4 (4.4)
Yes	0.87	1.6 (1.0)	0.0 (0.0)	1.2 (0.6)	6.1 (4.2)
Months since move					
Moved in last year	0.13	2.3 (1.1)	0.0 (0)	2.1 (0.7)	11.3 (7.9)
Moved over a year ago	0.34	1.2 (0.8)	0.0 (0.0)	1.0 (0.5)	6.2 (4.3)
Never moved	0.53	1.0 (0.8)	0.0 (0.0)	0.8 (0.5)	4.8 (3.6)
Patient demographics					
Perm res. male<45	0.19	1.4 (1.0)	0.0 (0.0)	1.2 (0.6)	6.9 (4.8)
Perm res. male≥45	0.16	1.0 (0.6)	0.0 (0)	0.9 (0.3)	4.8 (3.0)
Perm. res. female<45	0.30	1.5 (1.1)	0.0 (0.0)	1.2 (0.8)	7.1 (5.3)
Perm. res. female≥45	0.22	0.9 (0.6)	0.0 (0)	0.8 (0.4)	4.7 (3.2)
Temporary resident	0.13	1.4 (0.9)	0.0 (0.0)	1.2 (0.6)	6.9 (4.8)
Geographic location					
Rural	0.19	2.1 (0.4)	0.0 (0)	2.0 (0.3)	8.0 (3.6)
Suburban	0.36	1.0 (0.7)	0.0 (0.0)	0.8 (0.4)	5.9 (4.2)
Urban (Trondheim)	0.45	1.1 (1.0)	0.0 (0.0)	0.9 (0.6)	5.5 (4.4)

Notes: This table compares the perceived value of the mean (median) attentive patient's optimal GP choice under each mechanism, relative to the status quo mechanism Waitlists. Value is reported in a perpetuity equivalent monthly flow payoff in terms of travel time minutes to GP. Statistics are averaged across all attentive patients in months 392–451 of the simulation, holding each patient's current GP fixed at their assignment under Waitlists. See Appendix D.3 for additional details.

measure. Table 6 reports that patients whose current GP under Waitlists is undersubscribed under TTC have substantially lower choice values in the latter mechanism. The harm is significant in magnitude—the perpetuity equivalent of 0.8 minutes' travel time. Although these harms are offset by larger gains for patients with oversubscribed GPs, they may raise equity concerns for policymakers. In particular, it may seem unfair to disadvantage patients who already have a less desirable GP. This motivates exploring alternative mechanisms that attempt to improve outcomes for this group of patients.

Addressing Harms using DA and TTCP. The third column of Table 5 shows that DA achieves essentially no improvement over the status quo mechanism. Recall that DA does not execute trades that violate FCFS priority; patients may only swap GPs if all patients in front of them on their respective waitlists are also reassigned that month. As a result, DA reassigns few patients earlier than under Waitlists, and since those patients are already near the front of the queue, their waiting time reduction is small. This shows that there is a fundamental trade-off between eliminating envy—here, not allowing patients to "cut" in line—and finding additional gains from trade in this market.[41]

In contrast to DA, TTCP attempts to improve outcomes for patients with undersubscribed GPs without preventing feasible trades among patients with oversubscribed GPs. This is done by prioritizing patients with undersubscribed GPs above those with oversubscribed GPs, in effect ensuring that vacant slots are first offered to patients unlikely to benefit from cycles. The fourth column of Table 5 shows that TTCP achieves the majority of the welfare gains of TTC (0.46 vs 0.75 minutes in terms of the flow payoff measure; 1.08 vs 1.25 in terms of the NPV measure). Compared to TTC, more patients are standing on a waitlist at a given time, and patients expect to wait 1.1 months longer at their chosen GPs. For a patient with an oversubscribed GP, the expected waiting time for the average GP in their choice set rises from 10.7 to 12.2 months. However, this is offset by a much larger drop in expected waiting times for patients with undersubscribed GPs, from 22.8 to 18.9 months. These patients become more selective, requesting a GP with an average preference rank of 2.04 instead of 2.29, while patients with oversubscribed GPs become only slightly less selective.

One reason why TTCP yields smaller welfare gains than TTC is that prioritizing patients with undersubscribed GPs may reduce the total number of switches. When a patient with an undersubscribed GP is reassigned, the slot they vacate on their current GP's panel is unlikely to be demanded by another patient. Thus, the "chain" created by their assignment to a vacancy is short. Rather than facilitating more assignments, as would often occur under TTC, vacancy assignments usually end with the assigned patient. Indeed, the rate at which the waitlist moves from the front is 30 percent lower under TTCP than under TTC (c.f. Appendix Table D.2). Patients with undersubscribed GPs still benefit because they are placed higher on the waitlist, but the waitlists move more slowly overall. This illustrates the importance of explicitly considering the role of endowments, which determine the supply of GP slots, in our waitlist design counterfactuals.

Despite yielding smaller improvements than TTC, the results from TTCP are encouraging for policymakers concerned about the distributional consequences of introducing TTC in a market where some patients have more desirable endowments than others. The harms to patients with undersubscribed GPs are nearly eliminated—they enjoy equally valuable switching opportunities as under the status quo, according to the NPV measure in Table 6. At the same time, patients with oversubscribed GPs retain the benefits of being able to trade GPs, and

[41] The result that DA heavily restricts gains from trade mirrors empirical findings in other studies, e.g., Combe, Tercieux and Terrier (2022).

are still considerably better off than under the status quo.

V.E Discussion

We find considerable scope for improving Norway's GP allocation mechanism through simple changes to the assignment algorithm and priority rules. However, our proposals hold fixed many aspects of the mechanism as they are under Norway's current system, but which also could be redesigned. One important instance of this is that in all of our proposals, patients may only sit on one GP's waitlist at a time. This fundamentally limits the mechanism's ability to find gains from trade because patients may only express that one GP is preferred to their current assignment. Allowing patients to submit longer or even unrestricted rank order lists could dramatically reduce waiting times if patients would rank many GPs. This possibility raises an interesting trade-off between shorter waiting times and higher post-reassignment match quality. On one hand, asking patients to commit to one GP elicits cardinal information about preferences; desirable GPs are allocated to the patients most willing to wait for them. On the other hand, if there are long waiting times in equilibrium, patients spend more time with their mismatched current GP before reassignment.

Assessing the equilibrium implications of allowing patients to rank multiple GPs is challenging because such a change introduces considerable computational complexity into the patient's choice problem. In static implementations of DA or TTC, patients can do no better than truthfully reporting their ordinal preferences. In a dynamic implementation of these algorithms, "strategy-proofness" (as defined) would no longer hold. It may be optimal for a patient to truncate their true preference list to ensure that they receive a more desirable GP, even if they must wait longer. Calculating a patient's optimal truncation point requires calculating their continuation value from every possible truncated list. Even mechanisms which allow patients to rank a small number of GPs becomes combinatorially complex in a market with many alternatives—if joining two waitlists were allowed, there are $\binom{J}{2}$ possible rank-order lists. Further, these alternative designs would require modeling patients' beliefs about the joint distribution of waiting times across multiple waitlists, as well as modeling their strategic behavior under much more complex choice problems than those considered so far. These considerations limit our ability to predict *equilibrium* allocations under these alternatives.

Instead, we consider a benchmark simulation, **Truthful TTC**, in which patients truthfully report their full preference list to the mechanism, and the TTC algorithm is run each month.

This mechanism in effect allows patients to join an unlimited number of waitlists, but is implemented in such a way that we *assume* patients do so truthfully (i.e., join the waitlists for *all* GPs preferred to their current GP). The mechanism is formally described in Appendix D.4, and the results are reported in Appendix Table A.4.

Truthful TTC effectively eliminates waiting times—almost all patients can be reassigned to some preferred GP at the end of the month in which they join the mechanism. This result is driven by the fact that most patients prefer many GPs to their current one (often including at least one undersubscribed GP). The reduction in waiting times does come at the cost of a less preferred reassignment than under TTC. The average preference rank of a patient's reassigned GP is 2.72 under Truthful TTC, compared to 1.63 under TTC. It is nonetheless striking that even when patients cannot choose to wait longer for a more highly preferred GP, most are still reassigned to one of their top few choices. In fact, mean patient welfare is *higher* under Truthful TTC than under any of our proposed mechanisms.[42] Patients are 1.04 minutes better off relative to Waitlists, a larger improvement than under TTC. However, it is worth noting that these gains are concentrated among a minority of patients. Patients who are highly mismatched with their current GP may prefer a regime with less choice but much shorter waiting times, as Truthful TTC provides. In contrast, our primary proposals of TTC and TTCP modestly reduce waiting times and improve post-reassignment match quality, yielding more widely distributed gains.

These results suggest that it may be possible to dramatically reduce waiting times while still giving patients the ability to switch to GPs with whom they are quite happy. This could be implemented in a variety of ways, e.g., increasing the maximum number of GPs a patient can request (in other words, allowing patients to stand on multiple waiting lists), or allowing patients to request all GPs with certain characteristics (e.g., all GPs in their municipality of residence). However, such designs would add considerable complexity to a patient's decision problem when requesting to switch GPs, not to mention the researcher's equilibrium calculations. We therefore leave the equilibrium implications of such alternative mechanisms for future work.

[42]Note that we cannot evaluate welfare under Truthful TTC using the NPV measure because this measure relies on patients' equilibrium beliefs about waiting time. By construction, the Truthful TTC simulation does not require patients to form such beliefs, since they simply report their ordinal preference list truthfully.

VI Conclusion

This paper studies the market design problem of allowing patients to switch GPs in Norway's public healthcare system. We provide direct evidence of unrealized gains from trade under the current mechanism, suggesting scope for improvement. To predict outcomes under alternative mechanisms, we estimate a structural model of patient demand for GPs and apply the estimates within a dynamic equilibrium model of a GP reassignment system. Our results suggest that applying straightforward ideas from the market design literature—particularly the TTC algorithm—can reduce waiting times and improve patients' ability to obtain a well-matched GP. However, our results also highlight the fact that tools designed for static environments may have unintended consequences in a dynamic setting. In particular, introducing TTC within a dynamic reassignment system can leave some agents worse off relative to operating strictly first-come first-service wait lists (Norway's status quo mechanism). Patients who currently hold the least demanded GPs are especially disadvantaged. We then show that modifications to the priority system can improve outcomes for these patients while preserving about 60 percent of the gains from allowing patients to trade GPs.

We have focused on primary healthcare in Norway because the existence of formal waiting lists makes unrealized gains from trade visible to the researcher. However, similar gains could very well be present in other settings where where assignments are long-term and preferences may change over time, but where there is currently no formal way to register a desire to be re-assigned. Canonical market design applications such as public housing, specialized labor markets, public school seats, and hunting permits share many of the features of our present study, and often lack centralized reassignment mechanisms. Our findings yield both practical policy proposals and broader insights into how such a centralized mechanism might be best designed.

Even so, our analysis leaves several design aspects and broader questions unexplored. Regarding the design of the specific class of mechanisms we have considered, one could also optimize on dimensions we simply hold fixed. For example, it could be beneficial to adjust the frequency with which the matching algorithm is run, to modify the information made available to patients, or as discussed, to allow patients to join waiting lists for multiple GPs. More broadly, this study raises the question of how scarce resources should be rationed in dynamic assignment settings where prices do not clear the market. Many, if not most, settings of this type do not use formal waiting lists, and instead require agents to "check back later" for availability, effectively rationing desirable objects through agent effort rather than waiting times. Finally, to the extent that capacity constraints are somewhat flexible, there is a question of whether to impose hard caps (as in Norway), or to allow the market to clear through endogenous quality degradation through overcrowding (as in England). With some extension, our framework could be used to evaluate the choice between these various regimes.

References

Abaluck, Jason, and Abi Adams-Prassl. 2021. "What do Consumers Consider Before They Choose? Identification from Asymmetric Demand Responses*." *The Quarterly Journal of Economics*, 136(3): 1611–1663.

Abdulkadiroğlu, Atila, and Tayfun Sönmez. 1999. "House Allocation with Existing Tenants." *Journal of Economic Theory*, 88(2): 233–260.

Abdulkadiroğlu, Atila, and Tayfun Sönmez. 2003. "School Choice: A Mechanism Design Approach." *American Economic Review*, 93(3): 729–747.

Abdulkadiroğlu, Atila, Nikhil Agarwal, and Parag A. Pathak. 2017. "The Welfare Effects of Coordinated Assignment: Evidence from the New York City High School Match." *American Economic Review*, 107(12): 3635–89.

Agarwal, Nikhil. 2015. "An Empirical Model of the Medical Match." *American Economic Review*, 105(7): 1939–78.

Agarwal, Nikhil, and Paulo Somaini. 2018. "Demand Analysis Using Strategic Reports: An Application to a School Choice Mechanism." *Econometrica*, 86(2): 391–444.

Agarwal, Nikhil, Itai Ashlagi, Michael A. Rees, Paulo Somaini, and Daniel Waldinger. 2021. "Equilibrium Allocations Under Alternative Waitlist Designs: Evidence From Deceased Donor Kidneys." *Econometrica*, 89(1): 37–76.

Akbarpour, Mohammad, Eric Budish, Piotr Dworczak, and Scott Duke Kominers. 2023. "An Economic Framework for Vaccine Prioritization*." *The Quarterly Journal of Economics*, 139(1): 359–417.

Akbarpour, Mohammad, Shengwu Li, and Shayan Oveis Gharan. 2020. "Thickness and Information in Dynamic Matching Markets." *Journal of Political Economy*, 128(3): 783–815.

Arnosti, Nick, and Peng Shi. 2020. "Design of Lotteries and Wait-Lists for Affordable Housing Allocation." *Management Science*, 66(6): 2291–2307.

Ashlagi, Itai, Afshin Nikzad, and Philipp Strack. 2022. "Matching in Dynamic Imbalanced Markets." *The Review of Economic Studies*, 90(3): 1084–1124.

Baccara, Mariagiovanna, SangMok Lee, and Leeat Yariv. 2020. "Optimal dynamic matching." *Theoretical Economics*, 15(3): 1221–1278.

Bailey, Martha J., and Andrew Goodman-Bacon. 2015. "The War on Poverty's Experiment in Public Medicine: Community Health Centers and the Mortality of Older Americans." *American Economic Review*, 105(3): 1067–1104.

Baker, Laurence C, M. Kate Bundorf, and Anne Beeson Royalty. 2019. "The Effects of Multispecialty Group Practice on Health Care Spending and Use." National Bureau of Economic Research Working Paper 25915.

Bensnes, Simon, and Ingrid Huitfeldt. 2021. "Rumor has it: How do patients respond to patient-generated physician ratings?" *Journal of Health Economics*, 76: 102415.

Berry, Steven, James Levinsohn, and Ariel Pakes. 2004. "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: The New Car Market." *Journal of Political Economy*, 112(1): 68–105.

Berry, Steven T., and Philip A. Haile. 2014. "Identification in Differentiated Products Markets Using Market Level Data." *Econometrica*, 82(5): 1749–1797.

Bloch, Francis, and David Cantala. 2017. "Dynamic Assignment of Objects to Queuing Agents." *American Economic Journal: Microeconomics*, 9(1): 88–122.

Brown, Zach Y, Christopher Hansman, Jordan Keener, and Andre F Veiga. 2023. "Information and Disparities in Health Care Quality: Evidence from GP Choice in England." National Bureau of Economic Research Working Paper 31033.

Calsamiglia, Caterina, Chao Fu, and Maia Güell. 2020. "Structural Estimation of a Model of School Choices: The Boston Mechanism versus Its Alternatives." *Journal of Political Economy*, 128(2): 642–680.

Chartock, Benjamin L. 2021. "Quality Disclosure, Demand, and Congestion: Evidence from Physician Ratings." Bentley University Working Paper.

Che, Yeon-Koo, and Olivier Tercieux. 2023. "Optimal Queue Design."

Combe, Julien, Olivier Tercieux, and Camille Terrier. 2022. "The Design of Teacher Assignment: Theory and Evidence." *The Review of Economic Studies*, 89(6): 3154–3222.

Cutler, David, Robert Huckman, and Mary Landrum. 2004. "The Role of Information in Medical Markets: An Analysis of Publicly Reported Outcomes in Cardiac Surgery." *American Economic Review*, 94(2): 342–346.

Dahlstrand, Amanda. 2022. "Defying distance? The provision of services in the digital age." Centre for Economic Performance, LSE CEP Discussion Papers dp1889.

Ding, Hui, Yiwei Chen, Min Yu, Jieming Zhong, Ruying Hu, Xiangyu Chen, Chunmei Wang, Kaixu Xie, and Karen Eggleston. 2021. "The Effects of Chronic Disease Management in Primary Health Care: Evidence from Rural China." *Journal of Health Economics*, 80: 102539.

Empel, Giovanni, Hugh Gravelle, and Rita Santos. 2023. "Does quality affect choice of family physician? Evidence from patients changing general practice without changing address." *Economic Modelling*, 126: 106395.

Fack, Gabrielle, Julien Grenet, and Yinghua He. 2019. "Beyond Truth-Telling: Preference Estimation with Centralized School Choice and College Admissions." *American Economic Review*, 109(4): 1486–1529.

Fadlon, Itzik, and Jessica Van Parys. 2020. "Primary care physician practice styles and patient care: Evidence from physician exits in Medicare." *Journal of Health Economics*, 71: 102304.

Feigenbaum, Itai, Yash Kanoria, Irene Lo, and Jay Sethuraman. 2020. "Dynamic Matching in School Choice: Efficient Seat Reassignment After Late Cancellations." *Management Science*, 66(11): 5341–5361.

Gale, David, and Lloyd Shapley. 1962. "College admissions and the stability of marriage." *The American Mathematical Monthly*, 1(1): 9–14.

Gelman, A., J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. 2013. *Bayesian Data Analysis. Chapman & Hall/CRC Texts in Statistical Science*, CRC Press.

Godager, Geir. 2009. "Birds of a Feather Flock Together: A Study of Doctor-Patient Matching." University of Oslo, Health Economics Research Programme 2009-3.

Gravelle, Hugh, and Luigi Siciliani. 2008*a*. "Is Waiting-Time Prioritisation Welfare Improving?" *Health Economics*, 17(2): 167–184.

Gravelle, Hugh, and Luigi Siciliani. 2008*b*. "Optimal quality, waits and charges in health insurance." *Journal of Health Economics*, 27(3): 663–674.

Gruber, Jonathan, Thomas P. Hoe, and George Stoye. 2023. "Saving Lives by Tying Hands: The Unexpected Effects of Constraining Health Care Providers." *The Review of Economics and Statistics*, 105(1): 1–19.

Hansen, Elisabeth Holm, Erika Boman, and Lisbeth Fagerström. 2020. "Perception of the implementation of the nurse practitioner role in a Norwegian out-of-hours primary clinic: An email survey among healthcare professionals and patients." *Nordic Journal of Nursing Research*, 41: 54–60.

Heiss, Florian, Daniel McFadden, Joachim Winter, Amelie Wuppermann, and Bo Zhou. 2021. "Inattention and Switching Costs as Sources of Inertia in Medicare Part D." *American Economic Review*, 111(9): 2737–81.

Ho, Kate. 2006. "The Welfare Effects of Restricted Hospital Choice in the US Medical Care Market." *Journal of Applied Econometrics*, 21(7): 1039–1079.

Ho, Kate, Joseph Hogan, and Fiona Scott Morton. 2017. "The impact of consumer inattention on insurer pricing in the Medicare Part D program." *The RAND Journal of Economics*, 48(4): 877–905.

Hortacsu, Ali, Seyed Ali Madanizadeh, and Steven L. Puller. 2017. "Power to Choose? An Analysis of Consumer Inertia in the Residential Electricity Market." *American Economic Journal: Economic Policy*, 9(4): 192–226.

Kapor, Adam J., Christopher A. Neilson, and Seth D. Zimmerman. 2020. "Heterogeneous Beliefs and School Choice Mechanisms." *American Economic Review*, 110(5): 1274–1315.

Kapor, Adam J., Mohit Karnani, and Christopher A. Neilson. 2024. "Aftermarket Frictions and the Cost of Off-Platform Options in Centralized Assignment Mechanisms." *Journal of Political Economy*, 0(ja): null.

Kolstad, Jonathan T. 2013. "Information and quality when motivation is intrinsic: evidence from surgeon report cards." *American Economic Review*, 103: 2875–2910.

Larroucau, Tomas, and Ignacio Ríos. 2022. "Dynamic College Admissions." Working Paper.

Lee, Kwok Hao, Andrew Ferdowsian, and Luther Yap. 2024. "The Dynamic Allocation of Public Housing: Policy and Spillovers." Working Paper.

Leshno, Jacob D. 2022. "Dynamic Matching in Overloaded Waiting Lists." *American Economic Review*, 112(12): 3876–3910.

Lovdata. 2012. "Forskrift om fastlegeordning i kommunene, Helse- og omsorgsdepartementet." *Regulation; available at https://lovdata.no/dokument/SF/forskrift/2012-08-29-842*.

Mark, Nathaniel. 2021. "Access to Care in Equilibrium." *mimeo*.

McCulloch, Robert, and Peter E Rossi. 1994. "An exact likelihood analysis of the multinomial probit model." *Journal of Econometrics*, 64(1): 207–240.

Mora-García, C.A., M. Pesec, and A.M. Prado. 2023. "The effect of primary healthcare on mortality: Evidence from Costa Rica." *Journal of Health Economics*.

Narita, Yusuke. 2018. "Match or Mismatch? Learning and Inertia in School Choice." *SSRN Electronic Journal*.

Nichols, D., E. Smolensky, and T. N. Tideman. 1971. "Discrimination by Waiting Time in Merit Goods." *The American Economic Review*, 61(3): 312–323.

Pope, Devin. 2009. "Reacting to Rankings: Evidence from "America's Best Hospitals"." *Journal of Health Economics*, 28(6): 1154–1165.

Propper, Carol. 1990. "CONTINGENT VALUATION OF TIME SPENT ON NHS WAITING LISTS." *ECONOMIC JOURNAL*, 100: 193–199.

Propper, Carol. 1995. "The Disutility of Time Spent on the United Kingdom's National Health Service Waiting Lists." *The Journal of Human Resources*, 30(4): 677–700.

Robstad, Nastasja, Thomas Westergren, Eirin Mølland, Eirik Abildsnes, Kristin Haraldstad, Unni Mette Stamnes Köpp, Åshild Tellefsen Håland, and Liv Fegran. 2022. "Experiences of Norwegian child and school health nurses with the Starting Right child health assessment innovation: a qualitative interview study." *BMC Health Servives Research*, 22(278).

Roth, Alvin E. 1982. "The Economics of Matching: Stability and Incentives." *Mathematics of Operations Research*, 7(4): 617–628.

Roth, Alvin E., Tayfun Sönmez, and M. Utku Ünver. 2004. "Kidney Exchange*." *The Quarterly Journal of Economics*, 119(2): 457–488.

Santos, Rita, Hugh Gravelle, and Carol Propper. 2017. "Does Quality Affect Patients' Choice of Doctor? Evidence from England." *The Economic Journal*, 127(600): 445–494.

Shapley, Lloyd, and Herbert Scarf. 1974. "On cores and indivisibility." *Journal of Mathematical Economics*, 1(1): 23–37.

Shen, Yiwen, Carri Chan, Fanyin Zheng, and Gabriel Escobar. 2020. "Structural Estimation of Intertemporal Externalities on ICU Admission Decisions." *SSRN Electronic Journal*.

Su, Xuanming, and Stefanos Zenios. 2004. "Patient Choice in Kidney Allocation: The Role of the Queueing Discipline." *Manufacturing & Service Operations Management*, 6(4): 280–301.

Tay, Anthony. 2003. "Assessing Competition in Hospital Care Markets: The Importance of Accounting for Quality Differentiation." *The RAND Journal of Economics*, 34(4): 786–814.

The European Observatory On Health Systems and Policies. 2023. "Norway: Health system profile 2023."

Van der Vaart, Aad W. 2000. *Asymptotic statistics.* Vol. 3, Cambridge university press.

Verdier, Valentin, and Carson Reeling. 2021. "Welfare Effects of Dynamic Matching: An Empirical Analysis." *The Review of Economic Studies*, 89(2): 1008–1037.

Waldinger, Daniel. 2021. "Targeting In-Kind Transfers through Market Design: A Revealed Preference Analysis of Public Housing Allocation." *American Economic Review*, 111(8): 2660–96.

Wang, J., J. Hockenberry, S.Y. Chou, and M. Yang. 2011. "Do bad report cards have consequences? Impacts of publicly reported provider quality information on the CABG market in Pennsylvania." *Journal of Health Economics*, 30: 392–407.

Appendix A Data Details

A.1 Data Sources

The data for this study are derived from *Fastlegedatabasen*, maintained by the Norwegian Directorate of Health, as well as the administrative registries at Statistics Norway (SSB). Individuals in the data (both patients and GPs) are each assigned a unique identifier, which can be merged across datasets. Patients' municipalities of residence are identified at the monthly level using information from *Fastlegedatabasen*. Monthly municipality of residence is carefully tracked in this data because municipalities make monthly transfer payments to one another when a patient residing in one municipality enrolls with a GP located in another.

The raw GP enrollment and waitlist data is provided at the enrollment spell and waitlist spell level, respectively. Enrollment spells start and end only on the first and last days of a month (and so are in increments of full months). Waitlist spells can begin and end in continuous time. The time at which an individual joins a waitlist governs their priority in the waitlist. We convert the spells data to an individual-month panel. If an individual is on multiple waitlists in the course of a month, we take the most recent waitlist they were on. For the purposes of our primary analyses, we make only three restrictions to this data: (i) dropping individual-months that occur after the individual's registered date of death, (ii) dropping individuals under age 16, and (iii) dropping individual-months in which no current GP was registered. Appendix Table A.1 shows the number of individual-months dropped by these restrictions. Children under age 16 represent 17 of the data (and of the population).

Table A.1. Sample Construction Statistics

	2017		2018		2019	
Criteria	Number	Pct. of initial	Number	Pct. of initial	Number	Pct. of initial
Initial patient-months	63,437,631		63,813,290		64,212,740	
Registered after death	1,918	<0.01	48	<0.01	27	<0.01
Age < 16	11,475,019	0.18	10,694,796	0.17	9,918,551	0.15
No current GP	1,801	<0.01	2,066	<0.01	2,810	<0.01
Final total	51,958,893		53,116,380		54,291,352	

Notes: This table shows the number of patient-months each year dropped due to each sample selection criterion (in the order in which drops were made). The primary restriction on the data is to remove children under the age of 16.

A.2 Construction of Demand Estimation Sample

We estimate our model using the set of (adult) patient-months where the patient is resident in the Trondelag region and the month is between December 2018 and November 2019. Beyond this basic restriction, we must also make a number of further restrictions relating to the definition and construction of patients' GP choice sets. Our baseline choice set definition is a 60 minute drive time radius around patients' municipality of residence. Any individuals that were enrolled with (or joined a waitlist for) a GP further than 60 minutes driving time are dropped from the analysis. In addition, we drop patient-months in which the current or requested GP is exiting during the subsequent month. Finally, for patient-months in which no GP switch was requested, we drop observations where the patient was currently standing on a waitlist. These months reflect the outcome of a prior decision to join a waitlist, which we already capture in the set of patient-months where was a switch was requested.

Appendix Table A.2 summarizes this information, as well as the fraction of observations dropped at each step. We are left with 4,238,740 patient-months in which a switch was not requested, and 19,335 months in which a switch was requested. Both sets of observations together are informative for learning about when patients are attentive and consider switching GPs. Conditional on being attentive, the switcher-month observations are informative for learning about which GP characteristics patients value.

Table A.2. Demand Estimation Sample Construction

	Non-switches		Switches	
Criteria	Number	Pct. of initial	Number	Pct. of initial
Initial patient-months in Trondelag region	4,617,483		29,771	
Current GP further than 60 minutes	151,315	0.03	6,958	0.23
Requested a GP further than 60 minutes			2,976	0.10
Current GP exiting next month	158,457	0.03	404	0.01
Requested an exiting GP			98	<0.01
Currently on a waitlist	68,971	0.01		
Final total	4,238,740		19,335	

Notes: This table describes the sample selection criteria used in constructing the demand estimation sample. The unit of observation is a patient-month, and location of residence is measured at the monthly level. "Switches" are patient months in which the patient either joined a waitlist or else switched to a GP with open slots. "Non-switches" are patient months where a patient took no action. Both switches and non-switches are used in demand estimation.

Given the large size of the data, we proceed with estimation using all switcher patient-months and a random sample of 15,000 non-switcher patient-months. In estimation, we then re-weight the sampled non-switcher observations such that they represent the full set of observations. Given that patients with recent moves are disproportionately likely to be excluded from demand estimation because they hold a GP further than 60 minutes, we also adjust the sampling weights to match the original observed proportion of movers versus non-movers in both the non-switcher and switcher samples. Finally, once the random sample of non-switchers is drawn, we enforce a final restriction that all GPs remaining the analysis are chosen a sufficient number of times for a GP fixed effect to be estimated. We require each GP to be present in at least 400 individuals' choice sets. GPs that do not meet this requirement are dropped from all choices sets (and all patients that choose such a GP are also dropped). Our final estimation sample consists of 14,809 non-switcher months (re-weighted to represent 4,617,483) and 19,335 switcher months (re-weighted to represent 29,771).

Appendix B Additional Analysis

B.1 Mechanical Simulations: Implementation Details

Implementation of TTC. Only patients on waitlists participate in the TTC algorithm, as any patient not standing on waitlist retains their slot on their current GP panel. TTC is thus run to find a matching between (a) the set of patients standing on waitlists, and (b) the set of GP panel slots currently held by those patients. Preferences are all strict and are determined as follows:

(a) Patients standing on a waitlist first prefer their waitlist GP, then their current GP.

(b) GPs first prefer all their incumbent patients who are participating in TTC. Since these patients must be waiting on a waitlist, they have a global priority order determined by the moment in time at which they joined that waitlist. GPs prefer their incumbent patients in order of this global priority. GPs then prefer all the patients on their waitlist, in the order in which they joined the waitlist (which again corresponds to this global priority).

We iteratively look for cycles within chains that begin with a patient, starting with the patient who is highest priority on the global priority list and going down the global priority list from there.

Implementation of TTC on Historical Data. We implement the TTC algorithm monthly on the historical waitlists data between November 2016 (when waitlists were introduced) and December 2019 (the last month of our data). The purpose of this exercise is to generate a simple (but naive) estimate of how TTC would have changed the waiting lists had it been in place during this period. In that spirit, we hold all of the following objects fixed as they are observed in the data each month: patient entry (births or immigration), patient (deaths or emigration), administrative auto reassignments of patients, GP entry, GP exit, GP panel caps, and critically, patient actions. Patient actions each month are either (a) do nothing and remain with the current GP, (b) switch to an open GP, or (c) join a waitlist for a full GP. We then run the TTC algorithm on all patients standing on waitlists each month. Patients who can participate in a cycle are then reassigned their desired GP and removed from waitlists.

B.2 Preferences for GP Characteristics

This section investigates patient preferences for GP characteristics using a conditional logit analysis of GP choice. Given the size of the data, we limit attention to the geographic subsample of patients described in Section IV.A and Appendix A.2. Further, for the purposes of this specific analysis, we limit to the set of patient-months in which the patient requested to switch their GP, either by switching to an open GP or else by joining a waitlist. Within this subsample, we estimate the following conditional logistic regression specification:

$$u_{ijt} = -d_{ijt} + w_{ijt}\alpha + \delta_j + X_{ijt}\beta + \sigma_\epsilon \epsilon_{ijt}\,,$$

where d_{ijt} is driving time between patient i's municipality of residence and GP j's office, w_{ijt} is the number of patients on GP j's waitlist at the beginning of the month t when patient i made their switch request, δ_j is a set of GP fixed effects, X_{ijt} includes interactions between patient and GP age and gender, and ϵ_{ijt} is a type-1 extreme value idiosyncratic shock. We normalize the coefficient on driving time to -1.

Appendix Table B.1 reports results from three specifications. Column (1) excludes GP fixed effects and interactions between patient and GP characteristics. It controls only for GP age and gender, such that all horizontal GP differentiation comes from travel time and idiosyncratic taste shocks. Younger and female GPs are chosen more often, as are GPs with shorter waitlists. Column (2) adds observable patient-GP match-specific heterogeneity based on age and gender.[43] Compared to male patients, female patients have a strong preference for female GPs; they would be willing to travel about 5 minutes longer to see one. There is also clear, though weaker, homophily on age: all patient groups prefer younger GPs, but compared to a patient over age 45, a younger patient would travel about a minute longer to see a GP who is also under 45.

Table B.1. Preferences for GP Characteristics: Conditional Logit

	(1)		(2)		(3)	
Variable	β	SE	β	SE	β	SE
Travel time (minutes)†	−1.000		−1.000		−1.000	
Waitlist length	−0.026	0.002	−0.026	0.002	−0.090	0.006
Female GP	1.335	0.094				
× Temporary resident			2.490	0.247		
× Perm. res. female, age 16–45			3.723	0.173	1.209	0.274
× Perm. res. female, age 45+			2.838	0.189	0.131	0.286
× Perm. res. male, age 16–45			−1.910	0.221	−4.122	0.315
× Perm. res. male, age 45+			−2.461	0.258	−4.747	0.344
GP age 45+	−1.901	0.095				
× Temporary resident			−1.573	0.247		
× Perm. res. female, age 16–45			−2.353	0.168	−0.694	0.280
× Perm. res. female, age 45+			−1.544	0.183	0.060	0.290
× Perm. res. male, age 16–45			−2.118	0.217	−0.394	0.310
× Perm. res. male, age 45+			−1.343	0.243	0.303	0.329
Coeff. on epsilon shock	6.296	0.105	6.284	0.105	5.905	0.110
GP FE	N		N		Y	
Dep. variable mean	0.005		0.005		0.005	
# Observations	3,492,876		3,492,876		3,492,876	

Notes: This table reports coefficient estimates from conditional logistic regressions predicting the GP choices of patients who requested to switch GP between December 2018 and November 2019. The unit of observation is a (patient-month, GP) pair. The sample includes patient-months and GPs in Trondelag who meet the inclusion criteria for our structural estimation sample, with the additional restriction that each included GP is in at least 500 patient choice sets. The average patient-month has 192 GPs within choice set. Given this analysis conditions on switchers, we exclude each patient's current GP from choice set. The outcome variable is an indicator for whether the patient requested to switch to a specific GP. GP FE indicates that the specification includes a fixed effect for each GP, with one GP's fixed effect normalized to zero. In column (3), temporary residents serve as the omitted category for interactions between GP and patient characteristics. †By normalization

Column (3) adds GP fixed effects, absorbing any persistent GP-specific differences in desirability. The coefficients related to match specific heterogeneity—age and gender interactions

[43] Temporary residents are treated as a distinct demographic category because their demographic information is not available. We explored other GP and patient characteristics, and found few that were statistically and economically significant in explaining patients' GP choices.

and the standard deviation of the idiosyncratic shock—are very similar to column (2).[44] However, adding GP fixed effects triples the magnitude of the estimated coefficient on waitlist length. This is consistent with more desirable GPs having longer waitlists. Column (3) isolates responsiveness to variation in the length of *the same GP's* waitlist (relative to other waitlists) over time. This type of variation occurs naturally in queues due to statistical fluctuations in the number and types of agents and objects arriving over time (Waldinger, 2021; Leshno, 2022). The sensitivity of patients' choices to waiting time (information) will be a key determinant of the equilibrium implications of changing the design of the waitlist mechanism.

B.3 TTC is Not Pareto Improving: Styled Example

B.4 Predictors of Switching GPs

A number of patient characteristics appear predictive of a desire to switch GPs. For example, Table 1 shows that 34 percent of patients who had switched to an open GP and never used a waitlist over the period 2017–2019 had moved at some point during that period, compared only 6 percent of patients who had neither switched GP nor used a waitlist. To investigate the relative importance of various factors that may motivate GP switching, we regress an indicator for a GP switch request on patient characteristics. A GP switch request includes either an immediate switch to an open GP or a waitlist join. Our focal patient characteristics include time-invariant patient demographics and the timing of a switch relative to a move.

Appendix Table B.2 reports these results. The outcome variable (the switching indicator) is scaled by 100 for readability, so coefficients should be interpreted as percentage points. The overall probability of observing a switch request is 0.718 percent. Specification (1) includes only a set of five mutually exclusive and exhaustive patient demographic types. We find that temporary residents are substantially more likely to request to switch GPs than permanent residents, with a baseline probability of 1.501 percent. Among permanent residents, younger patients (particularly females) are more likely to switch.

[44] The levels of the coefficients are different than in column (2) because column (3) normalizes the preferences of temporary residents for GP age and gender to zero, but the *difference* between any two coefficients is nearly unchanged.

Table B.2. Predictors of Switching GPs

Variable	(1) β	(1) SE	(2) β	(2) SE
Patient demographic category				
Temporary resident	1.501	0.005	1.302	0.005
Perm. res. female, age 16–45	1.171	0.002	0.952	0.002
Perm. res. female, age 45+	0.501	0.001	0.459	0.001
Perm. res. male, age 16–45	0.782	0.001	0.568	0.001
Perm. res. male, age 45+	0.375	0.001	0.326	0.001
Timing relative to move				
Move in [t-6, t-1]; $<$30 min.			1.428	0.013
Move in [t-6, t-1]; $\geq$30 min.			3.243	0.013
Move in [t, t+1]; $<$30 min.			3.471	0.031
Move in [t, t+1]; $\geq$30 min.			11.451	0.037
Dep. variable mean	0.718		0.718	
R^2	0.009		0.021	
# Observations	154,793,455		154,793,455	

Notes: This table investigates observable predictors of requesting to switch GPs. The unit of observation is the patient-month. The sample includes all adult patients in Norway over the period January 2017 to November 2019. Regressions are linear probability models where the outcome variable is an indicator for a GP switch request (either a waitlist join or a switch to a GP panel with open slots). All covariates are indicator variables. For readability, the outcome variable is scaled by 100; coefficients should therefore be interpreted as percentage points.

Specification (2) introduces information on the timing of a given patient-month relative to a patient move. Conditional on a move being observed in the past 6 months, the current month, or the next month, we create four categories of moves based on an intersection of timing and distance of move. We find that switches are substantially more likely to occur in the month concurrent with or directly preceding the month of a move, and also that switches are far more likely to be associated with longer moves relative to shorter moves. For example, someone moving over 30 minutes away either this month or next month (i.e., in month [t, t+1]) has an extra 11.451 percent chance of requesting to switch GPs. Variation induced by moves therefore appears, unsurprisingly, to be highly important in determining when patients request to switch GPs.

B.5 Responsiveness to Aggregate Waitlist Length

An important question in our counterfactuals is whether more patients would request to switch GPs if expected waiting times fell systematically. Such a response would be predicted by a model where patients pay a "switching cost" at the time they *request* to switch GPs. In contrast, our proposed model of exogenous inattention rules out this type of response. As there is no direct switching cost, an attentive patient will always request to switch GP if any GP is preferable to their current one.

Testing for such a response is challenging because it requires aggregate variation in waitlist lengths, holding other demand and supply conditions fixed. One observation we can make is that the rates of switch requests have remained steady since waitlists were introduced in November 2016, and the number of waiting patients has grown almost linearly since then. If patients were more likely to make switch requests when waiting times were shorter, we would have expected a spike and then decline in switch requests after the introduction of waitlist. The time series evidence therefore weighs against a fully attentive model. However, it is also possible that patients gradually became aware of the new waitlist system, offsetting a deterring effect of increasing waiting times. We therefore also look for shorter-term variation in aggregate waitlist lengths.

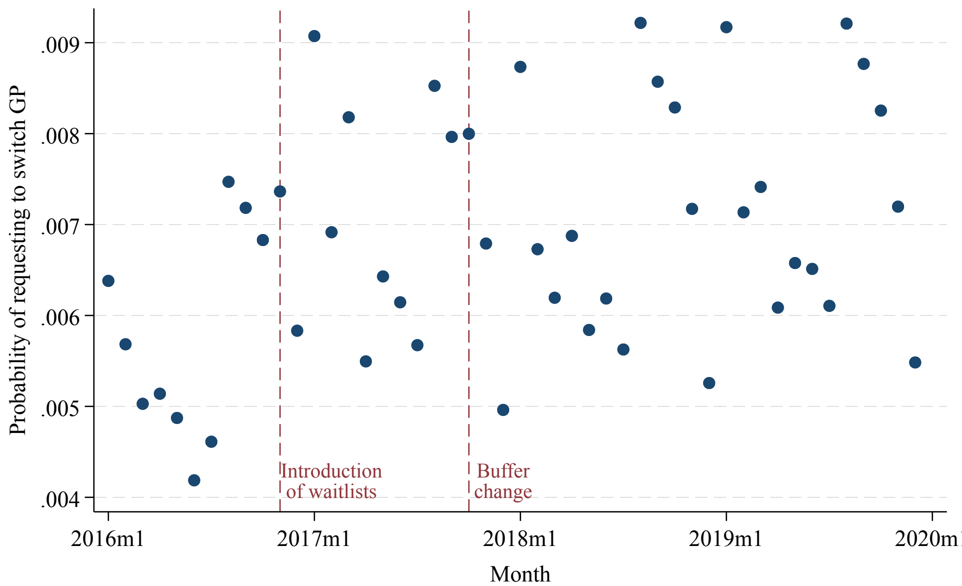
Another source of variation that occurred during our sample period—but before the period used for structural estimation—systematically reduced both perceived and actual waiting times. Until October 2017, Norway maintained a "buffer" of 20 slots on each GP's panel, and only assigned waiting patients after 20 slots were available.[45] The buffer was intended to prevent other additions to a GP's panel, such as births and administrative reassignments, from violating the panel cap. However, recognizing that this buffer kept patients waiting for GPs with open slots, the buffer was reduced from 20 to 10 slots in October 2017. GPs with 10-19 open slots had their panels filled with patients from the waitlist at the end of the month, resulting in a modest one-time drop in aggregate waitlist lengths. We use this variation to test whether aggregate switching rates increased immediately after this policy change, both overall and differentially by the amount different geographic regions were impacted.

Figure B.1 plots the monthly share of patients requesting to switch GP in a two-year window around the month of the buffer change. Monthly switching rates range between 0.5 percent and 1 percent, but there is no visible increase in the aggregate number of switch requests beginning

[45]The algorithm would fill all of the open slots once the buffer was exceeded, so a given GP did not always have 20 extra slots.

in October 2017. While there is not a clear aggregate response in switching requests, there was substantial variation across municipalities in the extent to which the change in the waitlist buffer affected waitlist lengths. In particular, smaller municipalities were much more affected. We next explore geographic heterogeneity in the impact of the buffer change.

Figure B.1. Probability of GP Switch Request: Event Study



Notes: The figure shows the average rate of GP switch requests in Norway over time, where a switch request includes both joining a waitlist and switching to an open GP. Waiters were introduced in November 2016. The waitlist buffer was reduced from 20 to 10 in October 2017.

The GPs directly impacted by the buffer change were those with 10–19 slots available prior to the change. For these GPs, all available slots were filled once the buffer was reduced to 10, but would not have been had the buffer remained at 20. We measure exposure to the buffer change in two ways. First, we calculate the fraction of GPs in each municipality who are near the buffer ("Near-Buffer"), meaning they had 10–19 open slots on their panel as of September 2017. In the median municipality, XX percent of GPs fit this description. Second, since waitlist lengths varied among Near-Buffer GPs, we multiply each municipality's Near-Buffer share by the average change in the length of the waitlist among Near-Buffer GPs. This measure isolates the average change in the length of *all* GP waitlists induced by the buffer change, which is more than 1.5 patients in the median kommune. We interact both measures with indicators for 1–3 and 4–12 months after the buffer change. Table B.3 reports these results. We find no differential impact of these exposure measures on switching rates, regardless of whether or not we control for a linear time trend in switching rates.

Table B.3. Difference-in-Differences: Switch Requests by Exposure to Buffer Change

	Near Buffer				Waitlist Change			
	(1)		(2)		(1)		(4)	
Variable	β	SE	β	SE	β	SE	β	SE
Oct.-Dec 2017 × 1 S.D. Exposure	0.120	0.153	0.121	0.153	0.000	0.117	−0.000	0.117
Jan.-Oct 2018 × 1 S.D. Exposure	0.153	0.135	0.154	0.135	−0.155	0.117	−0.155	0.117
Trend			−0.000	0.011			−0.000	0.011
Kommune FEs	Y		Y		Y		Y	
Month FEs	Y		N		Y		N	
Dep. variable mean	0.009		0.009		0.009		0.009	
# Observations	10,424		10,424		10,424		10,424	

Notes: The table reports estimates from linear regressions predicting the share of residents within a kommune-month who request to switch GP, weighted by population. The sample includes all kommune-months between October 2016 and October 2018. The outcome variable is the share of residents who requested during that month to switch to a new GP. In Columns (1) and (2), exposure is defined as the share of GPs "Near Buffer," i.e. with 10-19 open slots as of September 2017. In Columns (3) and (4), exposure is defined as "Waitlist Change," the change in waitlist length from September to October 2017 interacted with an indicator for GPs being near the buffer. Both measures are in standard-deviation units. All coefficients and standard errors are scaled by 1000 for readability.

Appendix C Estimation Details

This section provides details on the Gibbs' Sampler used to estimate the structural model parameters. We first describe the restrictions that the choice data place on the model primitives; then how we draw patient attention λ_{it} and flow payoffs v_{it}; and finally, how we update the discount factor ρ. The updating steps for (β, σ_ϵ) are standard. Unless otherwise specified, we condition on the month t in what follows and omit it from the notation.

Restrictions on Primitives. Consider a patient i with current GP j_0 who requests to switch in a given month. We learn two things from this patient. First, since they requested to switch GP, they were attentive: $\lambda_i = 1$. Second, their chosen GP j^* must have higher *expected net present value* than any other GP, given the patient's preferences and waiting time beliefs:

$$\underbrace{\mathbb{E}\left[e^{-\rho T_{ij^*}} \mid \mathbf{w}_i\right]}_{\omega_{j^*}(\rho)}(v_{ij^*} - v_{ij_0}) \geq \underbrace{\mathbb{E}\left[e^{-\rho T_{ij}} \mid \mathbf{w}_i\right]}_{\omega_j(\rho)}(v_{ij} - v_{ij_0}) \quad \forall j \in \mathcal{J}\,, \tag{5}$$

where we define $\omega_j(\rho)$ to be the expected discount factor to simplify notation. This set of inequalities implies a lower bound on the flow payoff v_{ij^*} from the chosen GP, and an upper bound on the flow payoff v_{ij} from each GP that was not chosen:

- **Upper Bound:** Rearranging Equation (5), for each $j \in \mathcal{J} \setminus j^*$,

$$v_{ij} \leq \underbrace{\frac{\omega_{j^*}(\rho)}{\omega_j(\rho)} v_{ij^*} - \frac{\omega_{j^*}(\rho) - \omega_j(\rho)}{\omega_j(\rho)} v_{ij_0}}_{ub_{ij}(v_{ij^*}, v_{ij_0}; \rho)},$$

where the notation $ub_{ij}(v_{ij^*}, v_{ij_0}; \rho)$ will be useful later.

- **Lower Bound** For chosen j^*,

$$v_{ij^*} \geq \underbrace{\max_{j \in \mathcal{J} \setminus j^*} \frac{\omega_j(\rho)}{\omega_{j^*}(\rho)} v_{ij} + \frac{\omega_{j^*}(\rho) - \omega_j(\rho)}{\omega_{j^*}(\rho)} v_{ij_0}}_{lb_{ij^*}(\mathbf{v}_i; \rho)}.$$

For patients who do not request to switch, it is not known whether they were paying attention. If the patient was not attentive ($\lambda_i = 0$), then not switching contains no information about their preferences. If the patient was attentive ($\lambda_i = 1$), then not switching implies that their current GP was preferable to all others: $v_{ij_0} \geq \max_{j \neq j_0} v_{ij}$.

Attention and Flow Payoffs. The Gibbs' sampler uses data augmentation to draw patients' attention and flow payoffs. The key challenge in drawing attention is calculating the likelihood that a non-switcher was attentive. By Bayes' Rule,

$$\begin{aligned} Pr(\lambda_i = 1 \mid \text{no switch}, X, j_0; \theta) &= \frac{Pr(\text{no switch} \mid \lambda_i = 1, X, j_0; \theta) Pr(\lambda_i = 1)}{Pr(\text{no switch} \mid X, j_0; \theta)} \\ &= \frac{Pr(\text{no switch} \mid \lambda_i = 1, X, j_0; \theta) Pr(\lambda_i = 1)}{Pr(\lambda_i = 0) + Pr(\text{no switch} \mid \lambda_i = 1, X, j_0; \theta) Pr(\lambda_i = 1)}, \end{aligned} \tag{6}$$

where $\theta = (\rho, \beta, \sigma_\epsilon(\cdot), p^\lambda(\cdot))$ collects the model parameters. By definition, $Pr(\lambda_i = 1) = p^\lambda(X_i)$ and $Pr(\lambda_i = 0) = 1 - p^\lambda(X_i)$. We can thus write

$$\begin{aligned} Pr(\text{no switch} \mid \lambda_i = 1, X, j_0; \theta) &= Pr(v_{ij_0} \geq \max_{j \neq j_0} v_{ij} \mid X; \theta) \\ &= \int_{-\infty}^{\infty} \Phi \left[\Pi_{j \neq j_0} \left(\frac{v_{ij_0} - X_{ij}\beta}{\sigma_\epsilon} \right) \right] \phi \left(\frac{v_{ij_0} - X_{ij}\beta}{\sigma_\epsilon} \right) dv_{ij_0}, \end{aligned} \tag{7}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal CDF and PDF, respectively. For each inattentive patient, the estimator evaluates this integral using Gauss-Hermite quadrature. Each patient's λ_i is an iid Bernoulli draw with probability defined in Equation 7 for non-switchers. The attention probabilities $p^\lambda(x)$ are then drawn from their posterior Beta distributions, given attention draws across all periods t:

$$p^{\lambda}(x) \mid \{\lambda_{it}\} \sim \text{Beta}\left(\alpha + \sum_{(i,t):X_{it}=x} \lambda_{it}\ ,\ \varphi + \sum_{(i,t):X_{it}=x} (1-\lambda_{it})\right), \tag{8}$$

where the prior is Beta(α, φ). In practice, we resample the individual attention draws and parameters every 10 iterations of the Gibbs' sampler.

Drawing patients' flow payoffs for each GP is more straightforward. For attentive patients, we redraw the flow payoff from each non-chosen GP from a truncated normal distribution with upper bound $ub_{ij}(v_{ij^*}, v_{ij_0}; \rho)$, and the flow payoff v_{ij^*} for the chosen GP from a truncated normal distribution with lower bound $lb_{ij^*}(\mathbf{v}_i; \rho)$.

Discount Factor. The discount factor is updated via a Metropolis-Hastings step which requires calculating the likelihood of the data given the other parameters. The algorithm begins with a previous draw ρ_{b-1} and a proposal distribution $F_\rho(. \mid \rho_{b-1}; \tau)$, which is truncated normal.[46] Each iteration, the following steps occur:

(i) Take a draw $\tilde{\rho}_b \sim F_\rho(. \mid \rho_{b-1})$

(ii) Calculate the ratio $r(\tilde{\rho}_b, \rho_{b-1}) = \frac{L(\tilde{\rho}_b)}{L(\rho_{b-1})}$, where $L(\rho)$ is the likelihood of the data given ρ and the other current draws of the model parameters.

(iii) Set $\rho_b = \left\{\begin{array}{lll} \tilde{\rho}_b & w.p. & \min\{1, r(\tilde{\rho}_b, \rho_{b-1})\} \\ \rho_{b-1} & w.p. & 1-\min\{1, r(\tilde{\rho}_b, \rho_{b-1})\} \end{array}\right\}$.

The main difficulty in implementing this is calculating the likelihood

$$L(\rho) = \Pi_{i:\lambda_i=1}\ Pr\left(j^* = \arg\max_{k \neq j_0}\ \mathbb{E}\left[e^{-\rho T_{ik}}\right](v_{ik} - v_{ij_0}) \mid \mathbf{X}_i, \mathbf{w}_i, \beta, \sigma_\epsilon, \rho\right) \tag{9}$$

for different values of the discount factor. This likelihood does not have a simple closed form, but it can be approximated with a high degree of accuracy using quadrature and exploiting the fact that the flow payoffs are conditionally independent given v_{ij_0}. Specifically, we can

[46] We experimented with an adaptive variance parameter τ, but settled on fixing $\tau^2 = 0.001$.

rewrite each probability in equation 9 as

$$L_i(\rho) = \int_{-\infty}^{+\infty} \int_{v_{ij_0}}^{\infty} Pr(v_{ik} \leq ub_{ik}(v_{ij}, v_{ij_0}; \rho) \;\; \forall k \neq j, j_0 \;\mid\; \beta, \sigma_\epsilon) \; dF_{ij}(v_{ij}) \; dF_{ij_0}(v_{ij_0})$$
$$= \int_{-\infty}^{+\infty} \int_{v_{ij_0}}^{\infty} \Pi_{k \neq j^*, j_0} \Phi \left(\frac{ub_{ik}(v_{ij^*}, v_{ij_0}; \rho) - X_{ik}\beta}{\sigma_\epsilon} \right) \; dF_{ij}(v_{ij}) \; dF_{ij_0}(v_{ij_0}), \tag{10}$$

where the last equality exploits conditional independence of v_{ik} given the flow payoffs from the patient's current and chosen GPs. To reduce the dimensionality of the integral, we condition on the value of v_{ij_0} when evaluating it, and evaluate the inner integral using Gauss-Laguerre quadrature.

Appendix D Counterfactual Simulation Details

D.1 Simulated Economy

As described in Section V.A, the simulated economy contains a finite set of patients I and GPs J. Patients have type $x \in \mathcal{X}$, where $\mathcal{X} = \{\text{demographic type}, \text{location of residence}\}$. GPs have type $z \in \mathcal{Z}$, where $\mathcal{Z} = \{\text{demographic type}, \text{office coordinates}, \text{panel cap}\}$.

GP characteristics are fixed over time. Patient characteristics evolve according to a stationary Markov process $M : \mathcal{X} \to \mathcal{X}$. Patients are assumed to indefinitely retain their gender ({Perm. res. female, Perm. res. male, Temp. res.}), so aging is the only relevant demographic transition. The transition process can therefore be thought of as drawing two independent shocks: a moving shock and an aging shock, where the probability of both shocks depends on a patient's current demographic type and location of residence. We calibrate transition probabilities to match observed transitions in Trondelag over 2017–2019. We then adjust the transition process to make it stationary.

Appendix Table D.1 reports details of the distribution of patient and GP types as well as the patient type transition process. Panel A describes patients. Seven percent of patients are temporary residents. Among permanent residents, half are female (with the remainder male), and 46 percent are age 16–45 (with the remainder 45 or older). Across all patients, the average probability of receiving an aging shock is 0.19 percent and the average probability of receiving a moving shock is 0.21 percent. The probability of aging is highest for temporary residents and lowest for females over age 45. The probability of moving is highest for female under age 45 and lowest for females over age 45. Conditional on moving, patients are more likely to move to a nearby location than a distant location. Given the observed distribution of patient and GP locations, the average patient has 94 GPs within 15 minutes driving time, and 190 GPs within 60 minutes driving time. But there is substantial heterogeneity. In some locations (rural areas), patient have only 5 GPs within 60 minutes, while in others (the central city Trondheim), there are 189 GPs within only 15 minutes.

Table D.1. Fundamentals of Simulated Economy

		Percentile				
	Mean	Min	25th	50th	75th	Max
Panel A: Patients (N = 371,536)						
Demographic Type						
Temporary resident	0.07					
Perm. res. female, age 16–45	0.21					
Perm. res. female, age 45+	0.26					
Perm. res. male, age 16–45	0.22					
Perm. res. male, age 45+	0.25					
Other Characteristics						
Prob. aging shock	0.002	0.001	0.002	0.002	0.002	0.007
Prob. moving shock	0.002	<0.001	0.001	0.002	0.002	0.012
Num. GPs within 15 min.	95	0	15	34	189	189
Num. GPs within 60 min.	190	5	73	262	262	279
Travel time to avg. GP	94	47	47	60	118	523
Panel B: GPs (N = 425)						
Demographic Type						
Female, age≤45	0.31					
Female, age>45	0.17					
Male, age≤45	0.25					
Male, age>45	0.27					
Other Characteristics						
Panel cap	930	80	794	936	1,098	1,695
Travel time to avg. municipality	210	97	100	135	281	746
Travel time to closest municipality	6	3	5	6	7	23
GP fixed effect	-1.0	-10.5	-1.3	-1.1	-0.7	5.6
Panel C: Locations (N = 45)						
Population	8,256	46	1,318	3,144	5,658	163,939
Pct. temp. res.	0.06	0.02	0.05	0.06	0.07	0.10
Pct. perm. res. female	0.47	0.41	0.46	0.46	0.47	0.54
Pct. perm. res. ≤45	0.37	0.21	0.33	0.36	0.40	0.52

Notes: The table describes the fundamentals of our simulated economy. These fundamentals are calibrated to match the Trondelag region as of December 2019. The 45 patient locations correspond to municipalities in the Trondelag region.

Panel B of Table D.1 describes the 425 GPs in the simulated economy. 48 percent of GPs are female, and 56 percent are under age 45. Among female GPs, however, 65 percent are under 45, reflecting growth of gender equity in this profession in recent decades. The average GP has a panel cap of 930 patients. The smallest GP has a cap of 80, while the largest has a cap of 1,695. For the purpose of our simulations, we do not use the panel caps observed in the data, but rather let them be determined based on the set of patients used in the simulation. For GPs that have waitlists as of December 2019, we set their panel cap equal to the observed number of patients that are enrolled with that GP and who reside in the Trondelag region. For GPs without waitlists, we set their panel cap equal to their observed enrollment among Trondelag patients times their observed ratio of enrollment to panel cap in the full data.

Finally, Panel C of Table D.1 describes the 45 possible locations where patients may live. The largest location is Trondheim municipality, with a population of 163,939 (nearly half the total population in the economy). The smallest location is Røyrvik municipality, with a population of only 46. The average location has 6 percent temporary residents, 47 percent female permanent residents, and 37 percent young permanent residents. Urban locations have a higher fraction of temporary residents as well as young permanent residents.

D.2 Algorithm to Compute Equilibrium

For each mechanism, we compute a stationary equilibrium in which patients' beliefs about waiting time are consistent the waiting times implied by patients' optimal decisions. We initialize the economy to have no patients standing on waitlists (i.e., no patients in the reassignment mechanism). We draw a sequence of patient types (demographic type, location of residence, identity of mother if reborn) and patient attention shocks for 500 periods (months) for all 371,536 patients in our simulation. These draws are held fixed across counterfactual mechanisms.

The equilibrium algorithm works as follows. We search for a fixed point between patients' belief parameters $\mathbf{b} = (\eta, \kappa, \kappa_{OS}, \chi_0, \chi_1)$ and their sample analogs within the simulated economy. The algorithm works as follows. Iteration q begins with a vector of belief parameters $\mathbf{b}^q$. The following steps then occur:

(1) The simulation is run for 500 periods. Each period, the steps described in Section V.A occur, with the caveat that all demographic transitions and attention shocks have been predetermined. Each period, attentive patients sequentially enter the reassignment

mechanism and consider all GPs in the economy, observing their current waitlist lengths. Patients form beliefs about waiting time according to $\mathbf{b}^q$ and then decide which GP to choose. If they choose their current GP or an open GP, they are reassigned immediately and exit the mechanism. If they choose a GP with a waitlist, they wait there until they can be successfully reassigned by the mechanism, they get another attention shock, or they die.

(2) The simulation provides data on the distribution of realized waiting times given the optimal actions implied by beliefs $\mathbf{b}^q$. We use this information to construct the sample analog of belief parameters, relying only on the last 250 periods of the simulation to allow the economy to converge to a stationary distribution.

(3) Beliefs are then updated as a convex combination of the initial and implied values: $\mathbf{b}^{q+1} = \lambda^q \mathbf{b}^q + (1 - \lambda^q)\mathbf{b}'$. The factor λ^q determines how quickly beliefs are updated.

Implied beliefs. Table D.2 reports equilibrium belief parameters. Under the status quo Waitlists mechanism, panel vacancies arrive at a rate of 0.5 percent chance per month per panel slot. Under the mechanisms with TTC, such natural vacancies arise less frequently (0.1 percent change per month per panel slot), since many incumbent patients who switch away no longer leave a vacant seat in their wake. Under DA, on the other hand, more vacancies arrive, since exchanges among patients at the front of waitlists can be processed.

While TTC lowers the panel vacancy rate, it raises the waitlist departure rate, as waiters may now participate in a cycle. Under TTC, all patients perceive this rate as 4.69 percent per waiter ahead of them per month. Under TTCP, patients with undersubscribed GPs perceive it to be only 0.76 percent, since the only patients ahead of them on waitlists are other patients with undersubscribed GPs, who have no chance of participating in a cycle. Under TTC, a patient with an oversubscribed GP who is 10th on the waitlist for a GP with a panel cap of 1,000 believes they will participate in a cycle with probability $0.132 = \exp(-4.9079 - 0.6275 \times \log(50/1000))$ per month. As their position-relative-to-panel-cap increases, this probability declines log-linearly. The same patient would expect to participate in a cycle with monthly probability 0.031 in waitlist position 100.

Appendix Figure A.5 provides a depiction of how our beliefs model translated into expected waiting times and discount factors across mechanisms. A patient with an oversubscribed GP believes that the expected waiting time (in months) for a GP with panel size 1,000 with a waitlist length of 100 would be 18.8 under Waitlists, 16.8 under TTC, 17.0 under TTCP, and

17.9 under DA. For a patient with an undersubscribed GP, the corresponding beliefs would be 18.8 under Waitlists, 39.2 under TTC, 97.0 under TTCP, and 17.9 under DA. Note, however, that because patients with an undersubscribed GP get waitlist priority under TTCP, they would rarely find themselves so far back on a waitlist. Appendix Figure A.6 reports the density of observed chosen waitlist lengths across mechanisms. Under TTCP, almost all the mass of chosen waitlist length is below a waitlist rank of 25. Finally, Appendix Figure A.7 provides a depiction of how our beliefs model interacts with panel capacities. Under the TTC mechanism, a patient with an oversubscribed GP believes that for a GP with a waitlist length of 50, expected waiting time would be 14 months if the GP had panel capacity of 750 and 11 months if capacity were 1,250. The average panel cap among GPs in our simulation is 930 (c.f. Appendix Table D.1).

Table D.2. Equilibrium Beliefs

	Waitlists	TTC	TTCP	DA
Panel vacancy rate (η)	0.0052	0.0010	0.0007	0.0052
Waitlist departure rate (κ)	0.0074	0.0468	0.0076	0.0074
Waitlist departure rate, curr. GP oversub. (κ_{OS})	–	–	0.0386	–
Cycle participation rate, curr. GP oversub. (χ_1)		-4.9074	-4.7997	
Cycle participation rate, curr. GP oversub. (χ_2)		-0.6273	-0.5816	

Notes: The table reports equilibrium belief parameters. In the TTCP mechanism, patients beliefs about the waitlist departure rate among waiters in front of them are κ for patients with an undersubscribed GP and $\kappa + \kappa_{OS}$ for patients with an oversubscribed GP. In all other mechanisms, patients have common beliefs about κ. For patients with an oversubscribed GP, the perceived probability of participating in a cycle in waitlist position s for a desired GP with panel cap N is given by $\chi = \exp(\chi_0 + \chi_1 \log(s/N))$.

D.3 Welfare Decomposition

A natural way to understand welfare under alternative mechanisms is to compare the NPV of an attentive patient's optimal choice of GP, which accounts for both the waitlist lengths they face when making a GP choice as well as their beliefs about how fast waitlists will move. Such a comparison is complicated, however, by the fact that at any given moment of attention, patients' *current GP* may change across mechanisms due to prior attention shocks the patient may have received in their lifetime. Since their degree of (dis)satisfaction with their current GP will directly affect how long they are willing to wait for a new GP, current GP is an important factor driving optimal choices. It is therefore useful to decompose NPV differences into (i) the component driven by a change in a patient's current GP, and (ii) the component driven by everything else, namely waitlist lengths, beliefs about how waitlist lengths map to wait times, and patients' optimal GP choice *conditional on current GP*.

To perform this decomposition, we introduce some additional notation. Throughout, we consider a single attentive patient in a single period whose preferences (and demographics) are held fixed across mechanisms. Express an attentive patient's perceived NPV from choosing GP j while currently enrolled with GP j_0 as

$$NPV(j; j_0, \mathbf{w}, \xi) = \mathbb{E}\left[\int_{\tau=0}^{T_j} e^{-\rho\tau} v_{j_0} d\tau + \int_{\tau=T_j}^{\infty} e^{-\rho\tau} v_{ij} d\tau \mid \mathbf{w}, \xi\right], \tag{11}$$

where $\mathbf{w}$ represents the vector of waitlist lengths for all GPs, ξ represents the patient's beliefs about the mapping from waitlist lengths to waiting time T_j, and v_j represents the per-period flow utility the patient derives from GP j. The NPV derived from a patient's optimal choice of GP is then given by

$$NPV^*(j_0, \mathbf{w}, \xi) = \max_{j \in \mathcal{J}} \quad NPV(j; j_0, \mathbf{w}, \xi), \tag{12}$$

where $\mathcal{J}$ is the set of all GPs. We can now isolate differences in the value derived from optimal behavior under each mechanism by differences in the arguments of NPV^*. As in Section V.D, we use the current Waitlists mechanism as our reference point. To that end, define the difference in the value derived from optimal behavior under focal mechanism M relative to the reference mechanism W as

$$\Delta NPV^{*,M-W} = NPV^*(j_0^M, \mathbf{w}^M, \xi^M) - NPV^*(j_0^W, \mathbf{w}^W, \xi^W), \tag{13}$$

where j_0^M is the patient's current GP under mechanism M, $\mathbf{w}^M$ is the vector of prevailing waitlist lengths under mechanism M, and ξ^M is the patient's (rational) beliefs about the mapping from waitlist lengths to wait times under mechanism M. Given this notation, we can then express the decomposition as

$$\begin{aligned}\Delta NPV^{*,M-W} = \quad & NPV^*(j_0^M, \mathbf{w}^M, \xi^M) - NPV^*(j_0^W, \mathbf{w}^M, \xi^M) && \text{(i)}\\ & + NPV^*(j_0^W, \mathbf{w}^M, \xi^M) - NPV^*(j_0^W, \mathbf{w}^W, \xi^W), && \text{(ii)}\end{aligned}$$

where as noted above, line (i) is the component of the welfare difference driven by a change in a patient's current GP, and line (ii) is the component driven by everything else, namely waitlist lengths, beliefs about how waitlist lengths map to wait times, and patients' optimal GP choice *conditional on current GP*. Because it more accurately represents the contemporaneous difference in value delivered by difference mechanisms, component (ii) is reported in the main text (Tables 5 and 6).

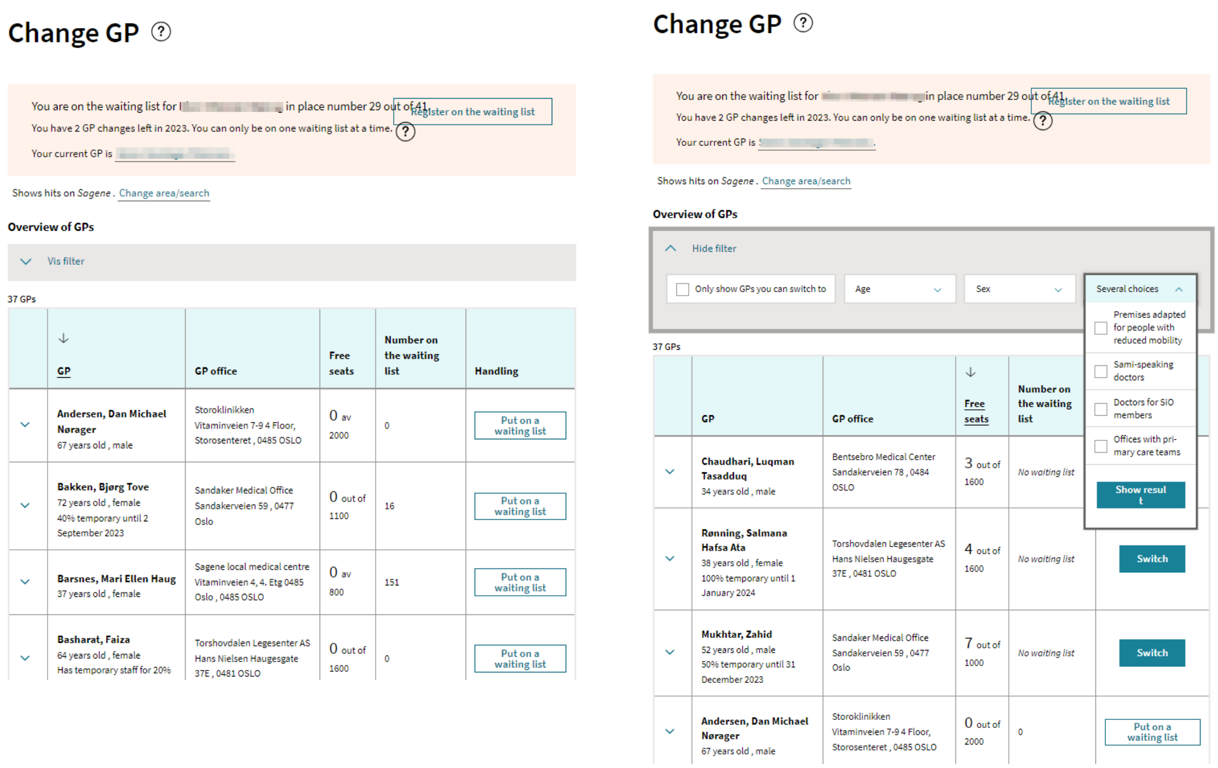
D.4 Benchmark Simulations

No Caps. Each GP's panel cap is infinite, so an attentive patient may switch to any GP immediately. This simulation provides an upper bound on the welfare that can be achieved by any mechanism with capacity constraints.

Truthful TTC. This mechanism is identical to TTC with the exception that patients can submit ROLs of arbitrary length. We run the simulation under the assumption that when they arrive to the mechanism, attentive patients truthfully report their full ordinal preferences over GPs, truncated at their current GP. This simulation provides a benchmark in which patients are "minimally selective," meaning they act as if they simply wish to switch to *any* preferable GP as soon as possible. Such behavior is not necessarily optimal for the patient—it may be preferable to truncate their reported ROL and thus wait a bit longer to receive a more highly-preferred GP—which is why this simulation is simply a benchmark.

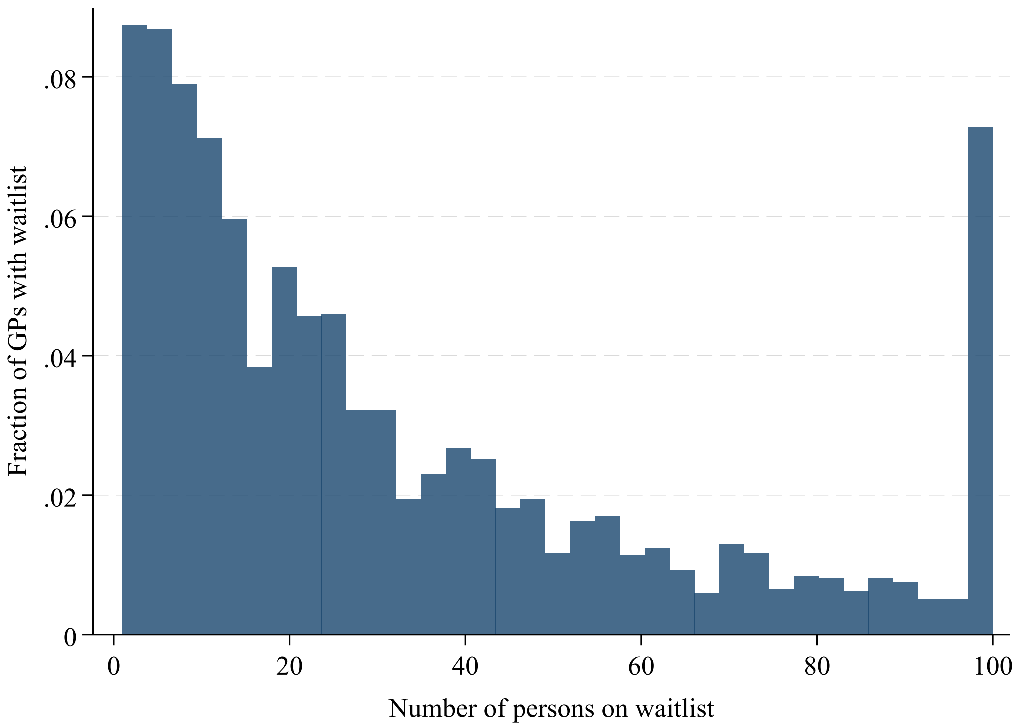
Figure A.1. HelseNorge Screenshots

(a) Sorted Alphabetically (b) Sorted by Free Seats



Notes: The figure shows two screenshots from the "Change GP" (*bytte fastlege*) tool on Norway's centralized online health platform, HelseNorge. The page shows the list of 37 GPs located in the Sagene neighborhood of Oslo. Panel (a) sorts this list alphabetically by GP last name (the default), and panel (b) sorts the list by the number of free slots available on each GP's panel. Only three GPs have available slots on their panel. (Webpage translated from Norwegian to English using Google Chrome, which slightly affects the rendering of graphics relative to the original webpage. Accessed August 18, 2023; *available at* https://tjenester.helsenorge.no/bytte-fastlege?fylke=03&kommuner=0301&bydeler=030103. This figure is referenced at footnote 8.)

Figure A.2. Distribution of Waitlist lengths in December 2019



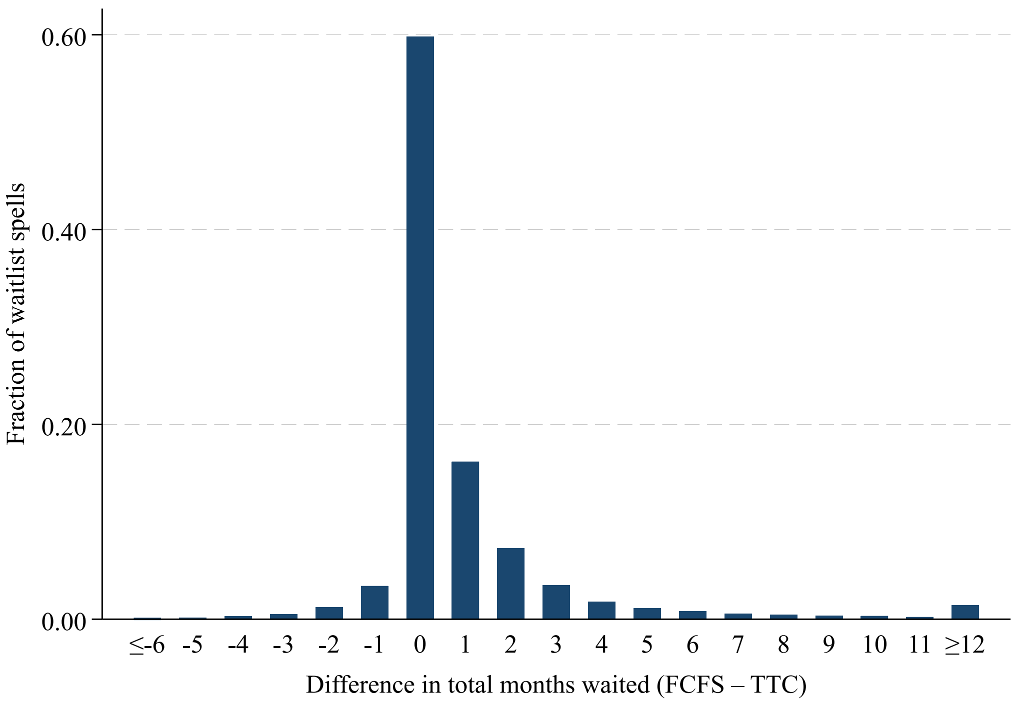
Notes: The figure shows the distribution of waiting list lengths in Norway in December 2019, among GPs that had a waiting list. Waitlist length is top-coded at 100 for readability. There were 3,695 unique GPs with waiting lists (out of 5,010 total GPs). This figure is referenced in footnote 16.

Table A.1. Patient Demographics by Outcome of Running TTC in December 2019

Sample demographic	Full Sample	Not on a waitlist	On a waitlist: Reassigned	On a waitlist: Not reassigned
Number of individuals	4,573,170	4,463,532	18,667	90,971
Pct. of individuals		0.98	0.00	0.02
Demographics				
Pct. female	0.50	0.49	0.66	0.66
Age	47	47	42	41
Years of education	13.1	13.1	13.4	13.3
Annual income (000 NOK)	413	414	397	373
Pct. temporary resident	0.07	0.07	0.04	0.12
Pct. ever moved	0.11	0.10	0.24	0.24
Choice of GP				
Pct. ever switched to open GP	0.13	0.13	0.17	0.30
Travel time to current GP (min.)	10.8	10.7	15.8	14.5
Pct. with GP of same gender	0.58	0.58	0.55	0.53
Use of waitlists				
Pct. ever on a waitlist	0.09	0.07	1.00	1.00
Number of months on a waitlist $\mid > 0$	6.4	4.9	7.9	10.7
Pct. waiting for GP of same gender	0.64	0.64	0.65	0.65
Travel time to wl. GP – curr. GP (min.)	-6.8	-7.2	-8.4	-5.6

Notes: The table provides summary statistics on adult patients based on the outcome of running TTC on waitlists as of December 2019. Summary statistics for each individual are calculated based on the time period 2017–2019, not just as they were observed in December 2019. "Ever" means at any point during 2017–2019. The first column reports means among all adult patients in the population. The remaining three columns are a partition of patients based on whether they were not standing on a waitlist in December 2019 and thus did not participate in TTC ("Not on a waitlist"), whether they were on a waitlist and were successfully reassigned by the TTC algorithm ("On a waitlist/Reassigned"), and finally those patients who were on a waitlist but were not successfully reassigned via TTC. This table is referenced in Section II.C.

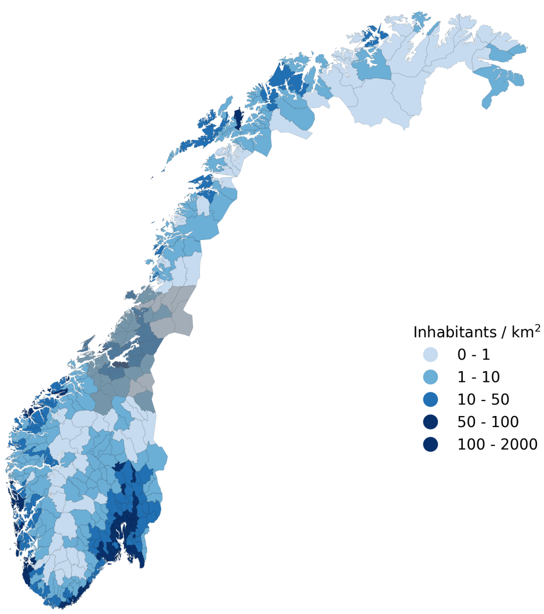
Figure A.3. Distribution of Waittime Differences



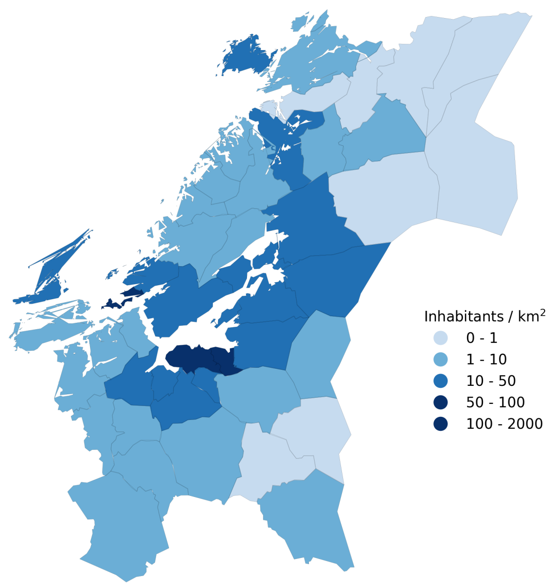
Notes: The figure shows the distribution of waiting time differences when running TTC monthly in the historical waitlist data. An observation is a waitlist spell. The figure reports the difference between the number of months the patient waited under the status quo first-come first-served waitlist mechanism (FCFS) and that under the waitlists plus TTC mechanism (TTC). While most patients wait for less time under TTC, 4.5 percent of patients wait for longer. This figure is referenced in Section II.C.

Figure A.4. Population Density Map

(a) All of Norway



(b) Trondelag Region



Notes: The figure shows a population density map of all of Norway as well as just the Trondelag region. In panel (a), the Trondelag region is shaded in gray. The outlined shapes within the map are municipalities. There are 421 municipalities in Norway and 58 in Trondelag (using 2019 region boundaries). The population center of Trondelag (the city Trondheim) lies at the center of the region in the darkest shaded municipality. This figure is referenced in footnote 28.

Table A.2. Comparison of Norway and Trondelag Region, 2019

Sample demographic	All of Norway	Trondelag
Panel A. Patient characteristics		
Number of individuals	4,633,395	412,774
Demographics		
Pct. female	0.50	0.49
Age	48	47
Pct. with post-secondary education	0.32	0.31
Annual income (000 NOK)	429	403
Pct. temporary resident	0.07	0.07
Pct. ever moved	0.04	0.06
Choice of GP		
Pct. ever switched to open GP	0.05	0.05
Travel time to current GP (min.)	10.7	10.9
Pct. with GP of same gender	0.58	0.57
Use of waitlists		
Pct. ever on a waitlist	0.05	0.05
Number of months on a waitlist $\mid > 0$	5.0	4.9
Pct. waiting for GP of same gender	0.64	0.65
Travel time to wl. GP – curr. GP (min.)	-6.7	-7.9
Panel B. GP characteristics		
Number of GP panels	5,549	474
Panel characteristics		
Enrollment cap	1,120	1,078
Pct. months with available slots	0.32	0.27
Pct. months with temporary GP	0.12	0.12
GP demographics		
Pct. female	0.43	0.46
Pct. rural	0.37	0.52
Age	47	45
Panel enrollment stats.		
Num. waiting on waitlist	24	22
Num. enrollees / cap	0.94	0.97

Notes: The table compares descriptive statistics between all of Norway and the Trondelag region in 2019. Panel A reports statistics on (adult) patients, and all values represent means over patient-months. "Ever" means at any point during 2019. Moves are counted only if they are across municipalities. Age, gender, education, and income data are not available for temporary residents, so those means are only among permanent residents. Panel B reports statistics on GPs, and all values in the table represent means over GP panel-months. GP enrollment and waitlist use statistics reflect the full population (including children under 16). This table is referenced in Section IV.A.

Table A.3. Summary Statistics on Realizations from Exogenous Processes

	Mean	SD
Number of patients	371,536	–
Number of attention shocks received	2,299	47
Number of moves	773	28
Number of aging shocks	735	26
Number of deaths	428	21
Number of moves in past 6 months	6,102	332
Number of attn. shocks in past 12 months	28,393	2,506

Notes: The table describes the realizations of exogenous processes in our simulated economy. These include the demographic transition processes (patients aging, dying, and moving) and the attention process (patients receiving attention shocks). Patients that die are immediately reborn, so there are a fixed number of patients in all periods. Statistics in the table are calculated across the 500 periods in the simulation. This table is referenced in section V.A.

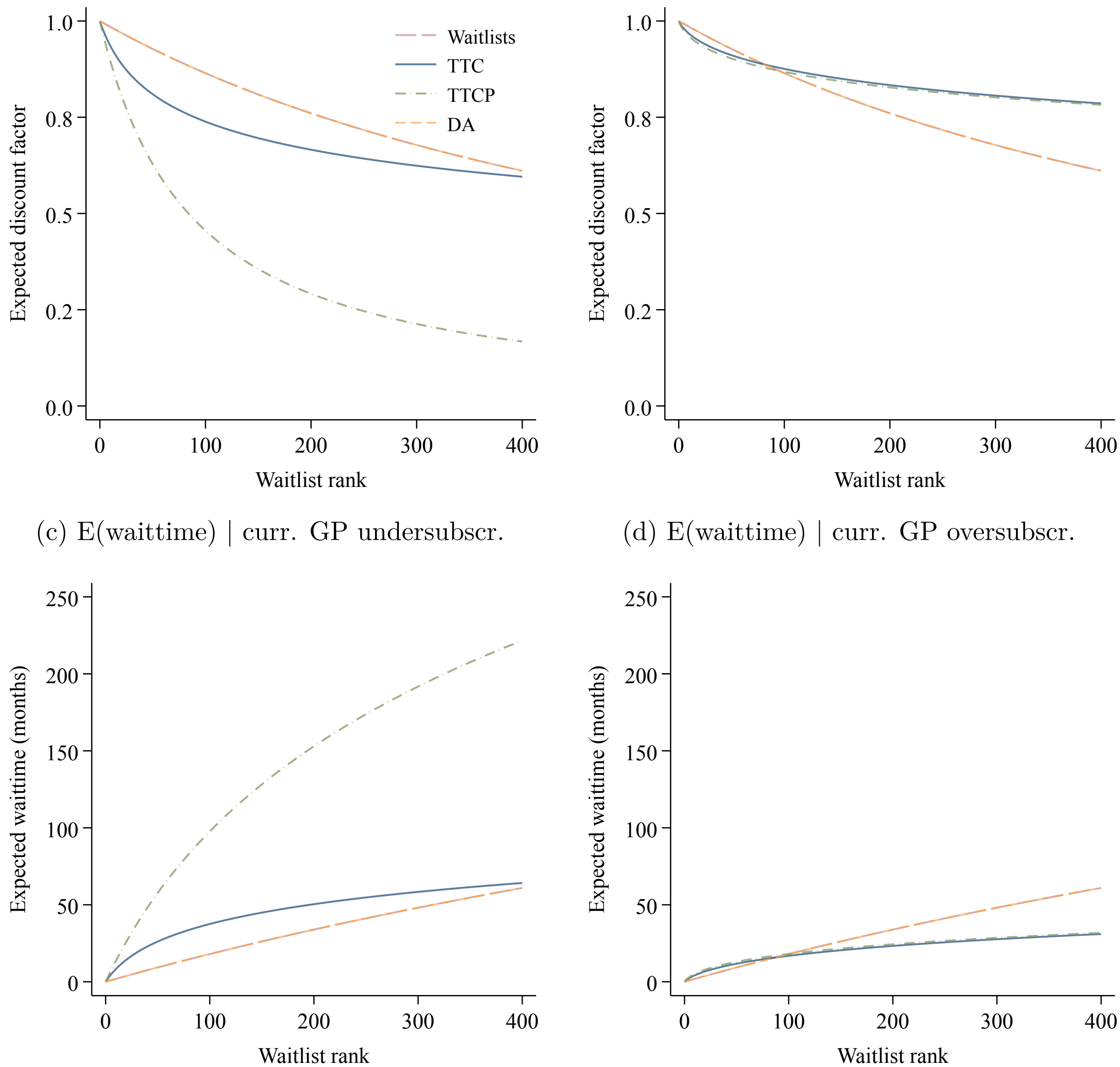
Table A.4. Results from Benchmark Simulations

	Waitlists	TTC	Truthful TTC	No Caps
GP waitlists				
Pct. of population on a waitlist	9.36	8.88	0.03	–
Pct. of GPs with a waitlist	82.2	78.3	32.3	–
Mean E(waittime) | curr. GP undersub.	16.7	22.8		–
| curr. GP oversub.	16.7	10.7		–
Attentive patient choices				
Mean E(waittime) at chosen GP	16.8	14.1		–
Pct. waitlist joins	85.2	84.6	93.6	–
| curr. GP undersub.	79.0	74.8	92.3	–
| curr. GP oversub.	86.2	86.3	94.1	–
True pref. rank of chosen GP	1.79	1.63	2.72	1.00
| curr. GP undersub.	1.94	2.29	2.67	1.00
| curr. GP oversub.	1.76	1.52	2.74	1.00
Realized assignments				
Travel time to current GP, mean (med.)	17.3 (6.5)	16.9 (6.4)	16.8 (6.5)	16.8 (6.4)
Pct. with same gender GP | young female	59.3	60.3	58.6	68.7
| young male	61.8	61.2	61.2	52.6
Welfare				
Flow payoff from current GP, mean (med.)	–†	0.75 (0)	1.04 (0)	5.34 (0)

Notes: The table reports results under the benchmark simulation of Truthful TTC (described in Section V.E). Results for Waitlists, TTC, and No Caps are reproduced from Table 5. Statistics are generated in months 392–451 of the simulation, out of 500 total months. They are first computed within month and then averaged across simulation months. E(waittime) is the expected waiting time implied by patient's equilibrium beliefs and current waitlist lengths. True pref. rank of requested GP is the rank of a patient's requested GP in their true flow payoff ordering. †By normalization. This table is referenced in Section V.E.

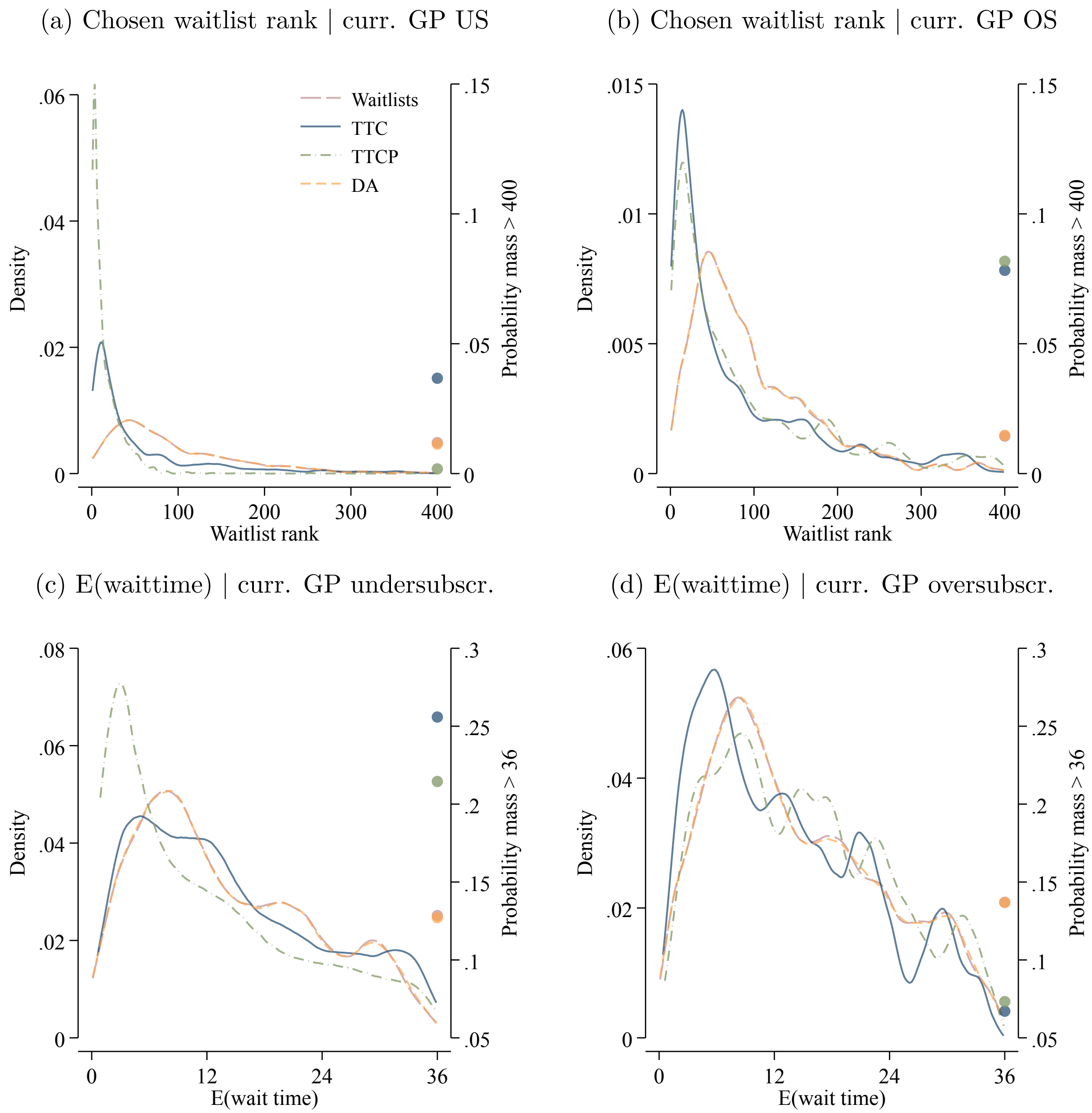
Figure A.5. Relationship Between Beliefs and Waitlist Rank by Mechanism

(a) EDF | curr. GP undersubscr.   (b) EDF | curr. GP oversubscr.



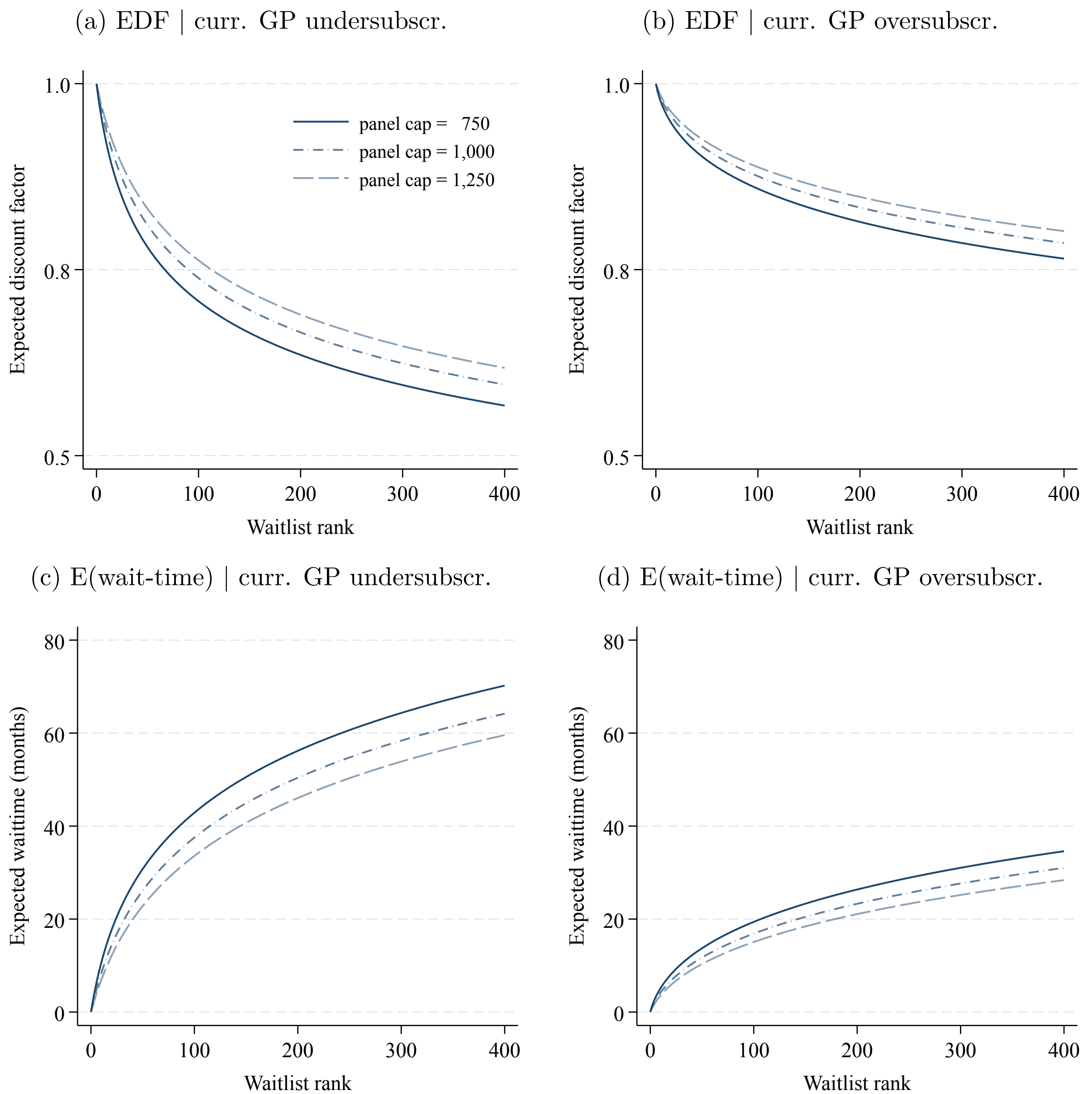
Notes: The figure shows the relationship between beliefs and waitlist rank across our four focal mechanisms, supposing all waitlists were for a GP with a panel cap of 1,000. Panels (a) and (b) report a patient's expected discount factor (EDF) as a function of waitlist rank. Panel (a) shows the EDF for a patient whose current GP is undersubscribed, while panel (b) shows the EDF for a patient whose current GP is oversubscribed. Panels (c) and (d) show the corresponding expected waiting times. This figure is referenced in Appendix D.2.

Figure A.6. Distribution of Chosen Waitlist Lengths by Mechanism



Notes: The figure shows the distribution of chosen waitlist ranks and the corresponding expected waiting times among attentive patients in each of our focal mechanisms. Panels (a) and (b) report the distribution of attentive patients' chosen waitlist rank conditional on being less than 400. The dots (scaled on the right axis) report the probability mass above this truncation point. Panels (c) and (d) report the distribution of corresponding expected waiting times, conditional on being below 36 months. Again, the dots (scaled on the right axis) report the probability mass above the truncation point. This figure is referenced in Appendix D.2.

Figure A.7. Relationship Between Beliefs and Waitlist Rank by Panel Cap (TTC Mechanism)



Notes: The figure shows the relationship between patient beliefs and waitlist rank for three different GP panel cap sizes, under the TTC mechanism. Panels (a) and (b) report a patient's expected discount factor (EDF) as a function of waitlist rank, if the waitlist considered was for a GP with a panel cap of 750, 1,000, or 1,250. Panel (a) shows the EDF for a patient whose current GP is undersubscribed, while panel (b) shows the EDF for a patient whose current GP is oversubscribed. Panels (c) and (d) show the corresponding expected waiting times. A higher panel capacity will make the waitlist move faster, and thus expected wait-time lower (and EDF higher). If a patient's current GP is oversubscribed, they will understand that they have the possibility of being reassigned via TTC, and thus have more optimistic expectations about waiting time. This figure is referenced in Appendix D.2.