American Economic Review 2026, 116(8): 3036–3075
https://doi.org/10.1257/aer.20240617

Designing Dynamic Reassignment Mechanisms: Evidence from GP Allocation†

By Ingrid Huitfeldt, Victoria Marone, and Daniel Waldinger*

We study the problem of designing a dynamic reassignment mechanism when agents' preferences over objects change over time. In the context of Norway's system for (*re*)*assigning patients to general practitioners* (*GPs*), *we provide direct evidence of misallocation under the current mechanism—patients waiting for each others' GPs*, *but who cannot trade—and estimate a structural model of GP-switching behavior to evaluate alternatives. Introducing top trading cycles* (*TTC*) *would*, *on average*, *reduce waiting times and increase patient welfare. However*, *patients endowed with less desirable GPs would be harmed. Prioritizing these patients can avoid these harms while preserving most of the gains from TTC.* (*JEL* D82, H51, I11, I13, I18)

Centralized (nonprice) assignment mechanisms are used to allocate many important resources in the economy, including schools, jobs, housing, and health care. 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 may change over time. Students may wish to transfer schools; public housing residents may want to down-/upsize 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*assigning agents to objects, when agents' preferences change over time. We make three conceptual and empirical

*Huitfeldt: Department of Economics, BI Norwegian Business School (email: ingrid.huitfeldt@bi.no); Marone: Department of Economics, Yale University (email: victoria.marone@yale.edu); Waldinger: Department of Economics, New York University (email: danielwaldinger@nyu.edu). Frank Verboven was the coeditor for this article. The original submission was handled by Katja Seim (Yale), who recused herself from subsequent revision rounds after Marone joined the Yale faculty. Author names in this article's byline are ordered alphabetically. We are grateful to Bob Town, Marika Cabral, Jacob Leshno, Nikhil Agarwal, Parag Pathak, Irene Lo, Oguzhan Celebi, Itai Ashlagi, Al Roth, Paulo Somaini, Adam Kapor, Maria Polyakova, Zarek Brot-Goldberg, Jose Ignacio Cuesta, Liran Einav, and workshop and seminar participants at UCL, NYU, Imperial College London, Penn State, Yale, Duke, Minnesota, Boston College, ASU, Berkeley, Cornell, Chicago Harris, Stanford, MIT Sloan, UT Austin, Northwestern, the NBER IO program meeting, and the University of Pennsylvania, as well as to the four anonymous referees for comments and suggestions that greatly benefited this research. We also thank Jon Georg Lund and Steinar Mathisen for providing valuable institutional details about the operations of the Norwegian GP Allocation System. Joridan Barash, Andrew Kim, Junrui Lin, and Kyle Han provided outstanding research assistance. This research received support from the Norwegian Research Council under Project 256678, which was approved by the Regional Committees for Medical and Health Research Ethics in Norway (REK reference 2017/373).
†Go to https://doi.org/10.1257/aer.20240617 to visit the article page for additional materials and author disclosure statement(s).

contributions. First, we provide direct empirical evidence of unrealized gains from trade in an important dynamic assignment market—the market for primary care physicians—henceforth, general practitioners (GPs)—in Norway. Second, we introduce an alternative mechanism that adapts a standard tool for centralized reassignment—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 and evaluate counterfactual mechanisms within a dynamic stochastic equilibrium model of a patient-GP (re)assignment 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 henceforth, patients) have free choice of GP. In practice, each GP has a cap on the number of patients they can have on their "panel," creating capacity constraints that limit patients' effective choice of GP.[1] In an effort to facilitate GP switching, in 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 GP's waitlist a time, and are then reassigned from the waitlist on a first-come, first-served basis to vacancies on the desired GP's panel.

The starting point for the paper is the observation that under this reassignment system, there may be patients waiting for each others' GPs, but who have no means by which to trade. Patients can only be reassigned to a vacant slot, precluding exchanges. These unrealized gains from trade naturally motivate the use of TTC, a centralized matching algorithm specifically designed to find mutually beneficial trades among individuals with endowed objects.[2] The algorithm looks for not only bilateral trades among pairs of patients, but also for "cycles" of trades among an arbitrary number of patients.[3] Applying TTC to our data, we find that 15 percent of patients on a waitlist in December 2019 (the last month of our data) could have been immediately reassigned through the algorithm. Moreover, 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.

Despite this clear evidence of unrealized gains from trade, the equilibrium consequences of incorporating TTC into a reassignment system of this type are not immediately obvious. Relative to a static environment, dynamics introduce two issues. First, TTC is not "strategy-proof" in a dynamic setting, meaning it is not necessarily optimal for a patient to request the GP they most prefer. Instead, patients' GP choices will likely respond to changes in waitlist lengths and in their beliefs about how quickly waitlists will move, both of which may vary across mechanisms.[4]

[1] Similar constraints exist in many primary care systems around the world, as well as in the context of health maintenance organizations (HMOs) in the United States.

[2] In a static environment, TTC also yields a Pareto improvement relative to the status quo allocation. That said, it does not necessarily have any optimality properties in a dynamic setting, where theoretical work has yet to fully characterize optimal mechanisms. See Section IV for a discussion.

[3] For example, suppose patient A is on the waitlist for the GP of patient B, who is on the waitlist for the GP of patient C, who is on the waitlist for the GP of patient A. A three-way exchange ("trading cycle") among the three patients can leave all with their requested GP.

[4] Our primary mechanisms of interest maintain the restriction that a patient may stand on at most one waitlist at a time. This amounts to limiting submitted rank-order lists to length two, in which case TTC is not strategy-proof

The gains from TTC will therefore depend not only on the value of resolving trades that exist at a given point in time, but also on patients' equilibrium responses to the change in the mechanism. Second, introducing TTC in a dynamic setting may not offer a Pareto improvement. Only patients with an oversubscribed GP, whose endowments are scarce resources, have the opportunity to trade. As a result, patients with undersubscribed GPs are effectively deprioritized and may experience systematically longer waiting times and lower welfare. Ultimately, the size of both the average gains from introducing TTC and any associated harms are an empirical question.

The rest of the paper studies the equilibrium impacts of introducing TTC and other related matching algorithms into Norway's waitlist system. Doing so requires economic modeling on two fronts. We first formulate a model of patient demand to switch GPs. A key input to the model is patients' beliefs about each GP's waiting time, which depend on the waitlist mechanism in use. We then introduce a dynamic stochastic equilibrium model of a patient-GP reassignment system. In the model, patients' desire to switch GPs arises stochastically through shocks to preferences and circumstances. The gains from introducing TTC are likely to be larger when (i) patients largely disagree on their desired GP and (ii) frequent changes in preferences and/or circumstances leave patients highly mismatched with their current GP. Furthermore, in a dynamic environment, patients' sensitivity to waiting time also impacts the value of such a reform through equilibrium waiting times and the value of the matches patients request. Our empirical analysis is designed to shed light on each of these elements in turn.

The demand model has two parts. First, each month, a subset of patients exogenously "pay attention" and actively consider requesting to switch GPs. This part of the model explains the fact that empirically, individuals rarely request to switch.[5] Second, "attentive" patients select their desired GP, taking into account both the flow utility they would derive from each GP as well as how long they expect to have to wait to be reassigned. The parameters we need to estimate include (i) the probability different patients are "attentive"—that is, make an active choice; (ii) the flow utility patients would derive from different GPs; and (iii) the discount rate, which governs sensitivity to waiting time.

We estimate the demand model parameters via a Gibbs sampler using monthly administrative data on Norway's GP assignment system. Our estimates imply substantial horizontal differentiation across GPs. Much of this differentiation is driven by geographic location, but we also find that patients have strong preferences for a doctor of the same gender (worth the equivalent of four to seven minutes of travel time) and similar age (one to two minutes). Moves are a particularly important driver of mismatch between patients and their GPs and strongly predict switch requests. Patients' preferences over GPs do not seem to be strongly influenced by their health, though patients with a chronic health condition do request to switch GP more often. Our preferred specification estimates an annual discount factor of approximately

even in a static setting. However, even if patients could submit unrestricted preference lists (as in the canonical TTC mechanism), they would have an incentive to truncate their submitted lists under TTC in the dynamic model we consider. See Section IVD for further discussion.

[5] Inertia or switching costs are common alternative explanations for infrequent switching behavior in repeated choice environments. Several descriptive patterns in the data weigh against these explanations here and motivate our focus on *exogenous* inattention. See Section II for further discussion.

0.91, consistent with descriptive evidence that patients' GP choices are responsive to waitlist lengths.

We apply these estimates within a dynamic stochastic equilibrium model of Norway's patient-GP reassignment system. The model is calibrated to match the basic elements of the Norwegian setting, including the distribution of patient and GP characteristics and the rate at which patients stochastically 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, 11 percent of the population would be standing on a waitlist in the stationary equilibrium, and the average patient would expect to wait over a year to switch to their chosen GP.

Our primary counterfactual is to run the TTC algorithm at the end of each month, after all naturally arising vacancies have been filled from waitlists. We find that relative to the status quo, the gains from introducing TTC are equivalent to reducing patients' travel time to their GP by 0.7 minutes for every patient in the economy. About a quarter of this improvement is directly due to patients' obtaining closer GPs, with the remainder due to better matching on other dimensions. These gains are economically meaningful, representing a welfare increase equivalent to what would be achieved by increasing every GP's capacity by approximately 2.5 percent under the status quo allocation mechanism. Overall, TTC benefits the majority of patients. Especially large benefits accrue to younger and female patients and recent movers, who are most likely to request to switch GPs and to use waitlists. Benefits accrue roughly evenly to sicker and healthier patients.[6]

As in our mechanical simulation, however, some patients are harmed. In particular, patients with undersubscribed GPs face longer waiting times and are worse off due to TTC. This harm is driven both directly by the fact that TTC prioritizes patients with desirable endowments, and indirectly by these patients' resulting 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, we implement the patient-proposing deferred acceptance (DA) algorithm instead of TTC. This mechanism strictly respects first-come, first-served waiting time priority, and thus does represent a Pareto improvement relative to the status quo.[7] However, it produces almost negligible gains, illustrating a fundamental trade-off in this setting between respecting first-come, first-served priority and exploiting gains from trade. Second, we implement a "TTC with priority" (TTCP) algorithm that prioritizes patients with undersubscribed GPs for panel vacancies.[8] The results are encouraging. TTCP achieves 56 percent of the welfare gains from TTC, while leaving patients with undersubscribed GPs just as well off as under the status quo.

[6] Among nonmovers, patients with a chronic health condition request to switch GPs more often than patients without and thus have a greater potential to benefit from TTC. However, chronic patients are also less likely to move, and these forces offset one another.

[7] 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.

[8] This idea is similar in spirit 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.

In a final analysis, we compare Norway's current mechanism to one in which there are not formal waitlists to ration excess demand for GPs. This case coincides with the regime in Norway before 2016 as well as that of most other primary care systems in which patients are formally assigned to GPs. Specifically, we simulate a mechanism in which patients may only choose from among GPs that have open slots at the moment they consider switching, meaning there is a substantial degree of "luck" involved. This exercise can be interpreted as a simulation of Norway's mechanism *before* the waitlists were introduced in 2016.[9] Strikingly, mean patient welfare is slightly *higher* than under Norway's current mechanism. However, median welfare is lower. The gains from not having waitlists are concentrated among a minority of patients who are highly mismatched with their current GP. These patients prefer an environment with more limited choice and no waiting times, whereas most patients prefer the ability to choose from a wider set of GPs. While suggestive only, these findings may partly explain why waitlists are rarely seen in other primary care systems. In contrast, the TTC and TTCP mechanisms reduce waiting times while also keeping the benefits from the increased choice that waitlists afford.

It is worth emphasizing that the normative standard of comparison used in this paper derives from patients' revealed preferences over GPs. Regulators may reasonably have alternative objectives, such as weighting health production beyond what patient choices reveal. Estimating how GP matches translate into health outcomes is beyond the scope of our analysis, and thus we are inherently limited in our ability to speak to objectives other than patient satisfaction. Nonetheless, we believe our analysis provides a first step in understanding how market design principles can inform the design of primary care systems.

Related Literature.—This paper contributes to a growing empirical literature on the design of centralized allocation mechanisms. In studying *dynamic reassignment*, we build on two specific strands of prior work. The first is *static reassignment*, in which all agents and objects are matched at a single point in time, and agents may have an endowed object. Canonical examples are the housing allocation problem with existing tenants and paired kidney exchange (Shapley and Scarf 1974; Abdulkadiroğlu and Sönmez 1999; Roth, Sönmez, and Ünver 2004). In this context, the TTC algorithm is known to be efficient and strategy-proof (ibid.).[10] These properties motivate adapting TTC to a dynamic environment, but as we will demonstrate, may break down when participants face dynamic incentives. The second is *dynamic assignment*, in which agents and objects arrive stochastically over time, but agents do not arrive with endowments. Examples include waitlists for public housing (Waldinger 2021; Lee, Ferdowsian, and Yap 2024), deceased donor kidneys (Agarwal et al. 2021), and hunting licenses (Verdier and Reeling 2022). Like in our setting, the trade-off for participants between shorter waiting times and a preferred assignment is central to

[9] Importantly, this simulation does not allow patients to "check back later" if their desired GP is not available when they first consider switching, which would add an element of patient effort to the rationing mechanism. We therefore view this simulation as simply a suggestive benchmark for what a "no waitlists" environment might look like in Norway. See Section IVD for further discussion.

[10] TTC has also been applied in settings of *static assignment* in which agents do not have endowments, particularly in the context of school choice (Abdulkadiroğlu and Sönmez 2003; Pathak and Sethuraman 2011; Leshno and Lo 2021). This literature highlights a tension between efficiently assigning students and respecting school priorities, which features prominently in our setting.

the optimal design of such mechanisms, as well as in revealed preference analysis of choice data.[11]

In the context of dynamic *re*assignment, Combe et al. (2022) study the reassignment system for French teachers in an infinite horizon environment, but model teachers as truthfully reporting their static preferences. Narita (2018) and Kapor, Karnani, and Neilson (2024) study models of school choice with aftermarkets in which agents can exhibit forward-looking behavior, but limit consideration to two periods. Our work builds on these studies by considering the combination of an infinite horizon and forward-looking agent behavior, which raises a new set of issues regarding how agents think about waiting times. Larroucau and Ríos (2022) also study a setting (college major choice) in which both these elements are present. Their focus is on learning and congestion externalities without the possibility of waitlists, while our focus is on finding gains from trade and the distributional issues that arise when agents may choose to wait. Waiting time is a natural market-clearing mechanism for processing reassignments in many settings (e.g., public housing, public schools, university parking spots), and our study represents a first step in understanding its equilibrium implications.

Finally, as in other countries, the existence of a national health insurance scheme in Norway motivates using a nonprice mechanism to allocate health care resources at the point of service. Our paper contributes to a literature studying the optimal design of such mechanisms and is among the first to bring the tools of market design to bear on this topic. The existing literature has largely focused on *service*-level allocation, in which context early work studied whether limiting capacity and running waitlists for nonemergency services could deter low-value care (Nichols, Smolensky, and Tideman 1971; Propper 1990, 1995; Gravelle and Siciliani 2008b). Subsequent work has studied the design of prioritization schemes on such waitlists (Gravelle and Siciliani 2008a; Shen, Chan, Zheng, and Escobar 2020; Gruber, Hoe, and Stoye 2023) as well as the relative merits of price versus waiting times as the rationing mechanism (Russo 2023). Our work, in contrast, considers the issue of *provider*-level allocation, where the potential for repeated interactions raises the issue of whether patients may wish to be reassigned. In a setting closely related to ours, Mark (2021) studies the Canadian primary care system, in which the process for switching GPs is decentralized and patients who wish to do so must exert costly effort to find vacancies. We see our work as complementary in that we study the design of a centralized mechanism, allowing us to apply tools from the field of market design.

The paper proceeds as follows. Section I introduces our setting and data and demonstrates the existence of gains from trade among waiting patients. Section II presents the structural model of a patient's choice of GP. Section III describes our estimation procedure and parameter estimates. Section IV presents our counterfactual simulations. Section V concludes.

[11] Several theoretical papers study how dynamic incentives impact the optimal design of dynamic assignment mechanisms (e.g., Su and Zenios 2004; Bloch and Cantala 2017; Arnosti and Shi 2020; Baccara, Lee, and Yariv 2020; Leshno 2022; Che and Tercieux 2023). To our knowledge, there is no theoretical characterization of optimal mechanisms for the class of dynamic reassignment models we consider. Existing theoretical work largely abstracts from strategic behavior, or focuses on a limited form of agent heterogeneity (Ashlagi, Nikzad, and Strack 2023; Akbarpour, Li, and Gharan 2020; Akbarpour, Budish, Dworczak, and Kominers 2024; Combe, Tercieux, and Terrier 2022). Work that has considered strategic behavior largely focuses on characterizing strategy-proof and stable mechanisms (Narita 2018; Feigenbaum, Kanoria, Lo, and Sethuraman 2020; Pereyra 2013).

I. Background, Data, and Descriptive Evidence

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. Health care utilization is primarily managed via supply-side forces. Central among these forces is a gatekeeping system whereby patients need a referral from a primary care provider before receiving specialist care. Such providers therefore play a central role in the health care system.

Primary care is almost exclusively provided by GPs.[12] 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.[13] Similar primary care systems exist in countries such as Canada, Great Britain, Italy, and Sweden, as well as in the context of health maintenance organizations (HMOs) in the United States. The key 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.[14] Similar to Medicare and Medicaid physician contracts in the United States, 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.[15] In 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 five working days, which in part motivates the existence of panel caps (Lovdata 2012).

Patients make GP enrollment elections via a nationally centralized online platform.[16] The system operates on a rolling basis, without any special enrollment periods or forced reenrollment decisions. Enrollment changes take effect on the first day

[12] 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). Primary care needs could include periodic well visits, nonurgent sick visits, obtaining prescriptions or referrals to specialist care, or receiving documentation for sick leave through the national sick leave scheme.

[13] 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.

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

[15] Operationally, local municipal governments work with GPs to set panel caps and negotiate any other idiosyncratic features of a GP's contract, such as reduced working hours or coverage for parental leave spells.

[16] The online platform is publicly available at https://tjenester.helsenorge.no/bytte-fastlege. Supplemental Appendix Figure E.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.

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.[17] Newborns are by default assigned to their mother's GP, regardless of whether the panel cap is violated. Every patient is thus assigned to a GP panel 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.[18] 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 as often as they would like.[19] Panel slots that become available are filled from the waitlist on a first-come, first-served basis, according to when 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.

B. *Data*

Our data are derived from two main sources (see Huitfeldt, Marone, and Waldinger 2026). First, we observe detailed administrative data on the GP assignment system itself (*Fastlegedatabasen*)(Norwegian Directorate of Health 2020a). 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 on health care utilization (Norwegian Directorate of Health 2020b; Norwegian Institute of Public Health 2020) and individual demographics, including age, gender, family relationships, income, education, and monthly municipality of residence (Statistics Norway (SSB)). Further details are provided in Supplemental Appendix A.1. We have data for the years 2014–2019.[20]

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 31,856 patients (0.6 percent) switched their GP each month. Once waitlists were introduced in November 2016, an average of 28 percent of switches were executed via a waitlist, whereas the remaining 72 percent were

[17] 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.

[18] In fact, a cottage industry arose of private companies that automated the process of monitoring the central web platform for when GP panel slots became available and selling an email/text alert service to consumers for a monthly fee.

[19] The centralized web platform provides real-time information on the number of patients on each waitlist, and if a patient is logged in, on their current position on the waitlist they are standing on (see Supplemental Appendix Figure E.1).

[20] An exception is health care utilization data, which we only have for 2014–2017.

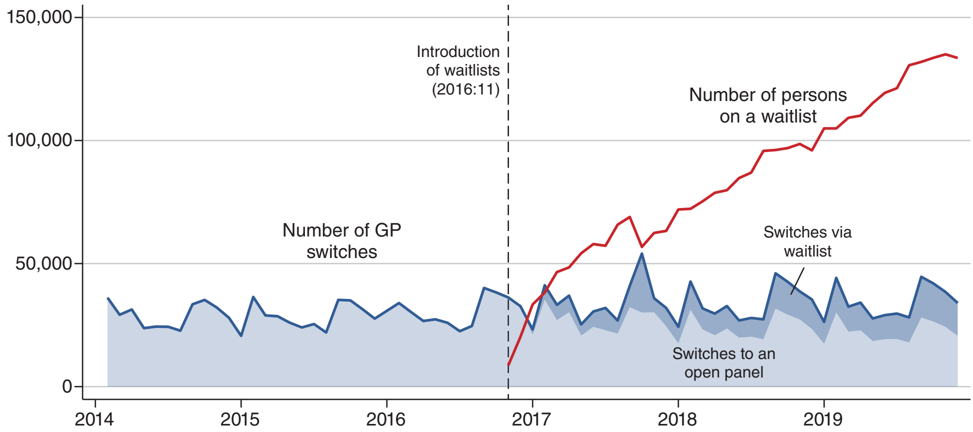

Figure 1. Number of GP Switches and Waitlist Use over Time

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).

switches to an open GP panel. The total number of switches per month did not change dramatically, but 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 patients). As of February 2024, this number had risen to 326,506 (6.5 percent of all patients).

Patients.—Our primary restriction on patients is to exclude those who are under 16 years old. Children's interactions with the health care 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 substantially complicate our analysis, 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. Patients are on average 47 years old and earn kr425,374 annually. Just over 7 percent of individuals are temporary residents, 32 percent have postsecondary education, 32 percent have a chronic health condition, and 10 percent moved to a new municipality at least once during 2017–2019.[21]

[21] Demographic characteristics such as age, gender, education, income, and presence of chronic health condition are only available for permanent residents. Where means of these variables are reported, they thus condition on the individual being a permanent resident. Monthly municipality of residence is available for all individuals. Supplemental Appendix A.1 provides additional details on the construction of these variables.

Table 1—Patient Summary Statistics

		Never used waitlist		
Sample demographic	Full sample	Never switched	Ever switched	Ever used waitlist
Number of individuals	4,780,647	3,875,753	497,111	407,783
Percent of individuals[a]		0.81	0.10	0.09
Demographics				
Percent female	0.50	0.48	0.51	0.64
Age	47	49	40	42
Percent with postsecondary education	0.32	0.31	0.32	0.37
Annual income (kr thousands)	425	437	358	390
Percent with chronic condition	0.32	0.32	0.30	0.34
Percent temporary resident	0.07	0.07	0.08	0.08
Percent rural	0.30	0.31	0.30	0.28
Percent ever moved	0.10	0.06	0.34	0.26
Choice of GP				
Percent 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
Percent with GP of same gender	0.58	0.57	0.59	0.56
Use of waitlists				
Percent ever on a waitlist	0.09	—	—	1.00
Number of months on a waitlist $\mid > 0$	6.5			6.5
Percent waiting for GP of same gender	0.64			0.64
Travel time to waitlist GP − curr. GP (min.)	−6.8			−6.8

Notes: The table provides summary statistics on adult patients present in the data from 2017 to 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, health, and income data are not available for temporary residents; these means therefore represent only permanent residents.

[a] For readability, all percentages are reported as proportions (e.g., 0.20 = 20 percent).

In terms of GP choice, 19 percent of patients requested to switch their GP at least once, and 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.[22] 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. Second, patients who requested to switch GPs (either via an open panel or a waitlist) are seven to nine years younger on average than patients who never requested to switch. They are not, however, commensurately less likely to have a chronic condition, suggesting that switching patients have overall worse age-adjusted health

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

Table 2—GP Panel Summary Statistics

	All	Undersubscribed	Oversubscribed
Number of GP panels	6,470	3,532	2,938
Percent of GP panels[a]		0.55	0.45
Panel characteristics			
Enrollment cap	1,138	1,143	1,135
Percent months with available slots	0.36	0.70	0.06
Percent rural	0.37	0.44	0.30
GP demographics			
Age	47	47	48
Percent months with female GP	0.42	0.34	0.50
Percent months with temporary GP	0.11	0.15	0.08
Panel enrollment stats.			
Number enrollees on a waitlist	18	21	14
Number waiting on waitlist	18	4	30
Number 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 to 2019. Except where otherwise specified, all values in the table represent means over GP panel-months. Oversubscribed GP panels are those that are at capacity for more than 75 percent of months. Enrollment and waitlist-use statistics reflect the full population, including children under 16.

[a] For readability, all percentages are reported as proportions (e.g., 0.20 = 20 percent).

status. Waiters are also disproportionately female, which could be driven by a scarcity of female GPs and/or a stronger gender homophily preference among female patients.

Finally, 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. Among those who switched GP but never used a waitlist—meaning they switched to an open GP—34 percent of individuals moved. Among those who used a waitlist, 26 percent moved. 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—that is, more likely to simply settle for an open GP. These facts suggest that geographic proximity is an important factor in GP choice. Indeed, the last row of the table shows that 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).

GPs.—Table 2 provides summary statistics on the 6,470 GP panels active during 2017–2019. As with patients, the data consist of an unbalanced panel at the GP panel-month level.[23] 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.

[23] 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.

The first column of Table 2 describes the full set of GP panels. The average panel had a cap of 1,138 patients 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.[24] The average GP panel-month had 18 people standing on its waitlist, and the average enrollment-to-cap ratio was 94 percent.

The remaining two columns of Table 2 separate GP panels into two subgroups: (i) those that were not consistently oversubscribed and (ii) those that were. 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 to 2019. There are several notable patterns. First, GP panels in urban areas and those served by female GPs are more likely to be oversubscribed, while GP panels served 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 overdemanded GPs. Finally, capacity utilization in the system is high. Even for consistently undersubscribed GPs, 87 percent of panel slots were occupied.

C. *Unrealized Gains from Trade*

Stylized Example of TTC.—Consider a hypothetical example of three GPs (A, B, and C), each of whom has a waiting list. Though each GP's panel is currently full, slots will periodically become available as patients on each panel die or switch to other GPs' panels. Under Norway's current system, waiting patients are assigned to these open slots on a strictly first-come, first-served basis. Trades among patients standing on waitlists are not permitted.[25] At any given time, however, there may be patients waiting for one another's GPs. Such patients could in principle just switch with one another, obtaining their desired GPs sooner while still respecting capacity constraints.

The purpose of the TTC algorithm is simply to take this idea further. Suppose there is a patient who currently has GP A but who is waiting for GP B, a patient who currently has GP B but is waiting for GP C, and a patient who currently has GP C but is waiting for GP A. A three-way exchange ("trading cycle") among the three patients can leave all with their desired GP, just as in the bilateral case. Supplemental Appendix Figure E.3 provides a depiction. In the three-way cycle, no pair of patients exchanges GPs, but each patient can be successfully reassigned while another patient in the cycle takes their former slot.

In general, there could be many such cycles in the GP waitlist system at any given time. And moreover, a given patient could have several different feasible trading

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

[25] Even when two people are each at the front of the waitlist for one another'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.

partners through different cycles that exist simultaneously. The TTC algorithm systematically finds and executes trading cycles of the kind described above, with no prescribed limit on the number of patients that could participate in a given cycle (the longest cycle we find involves 342 patients). At each iteration, the algorithm finds a cycle of patients and their GP slots, reassigns those patients to their requested GP, and removes them and their GP slots for the remainder of the algorithm (Shapley and Scarf 1974; Abdulkadiroğlu and Sönmez 1999). It repeats this process until there are no more cycles, meaning there are no remaining patients on GP waitlists who would wish to trade with one another. Patients who are not successfully reassigned through TTC retain their current GP and keep their current waitlist position.

Double Coincidence of Wants at a Point in Time.—We use the TTC algorithm to provide 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 developed in the remainder of the paper. We begin by running TTC at a single point in time: the last month of our data, December 2019. 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.[26]

The TTC algorithm takes two primary sets of inputs in our setting: (i) the set of participating patients and their rank-order preference lists over GPs and (ii) the set of participating GPs and their rank-order priority lists over patients. 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. GPs first prioritize all patients currently on their panel, guaranteeing each patient an assignment no worse than their current GP. GPs then prioritize the patients on their waitlist in descending order of waiting time. Supplemental Appendix B.1 describes the exact algorithm we use to find and clear cycles, as well as other implementation details. We run the algorithm and find that 20,377 people (15 percent of waiters) could have been immediately reassigned.[27] Supplemental Appendix Table E.1 describes the demographics of these patients (among adults only). Successfully reassigned patients tend 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.

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. Supplemental Appendix B.2 provides a detailed description of these results. We find that indeed, the most important source of horizontal preference heterogeneity is patients' geographic location relative to GP offices. We also find that patients have a preference for a same-gender and similar-age GP.

[26] Supplemental Appendix Figure E.2 provides a histogram of waitlist lengths across these 3,695 waitlists.

[27] 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).

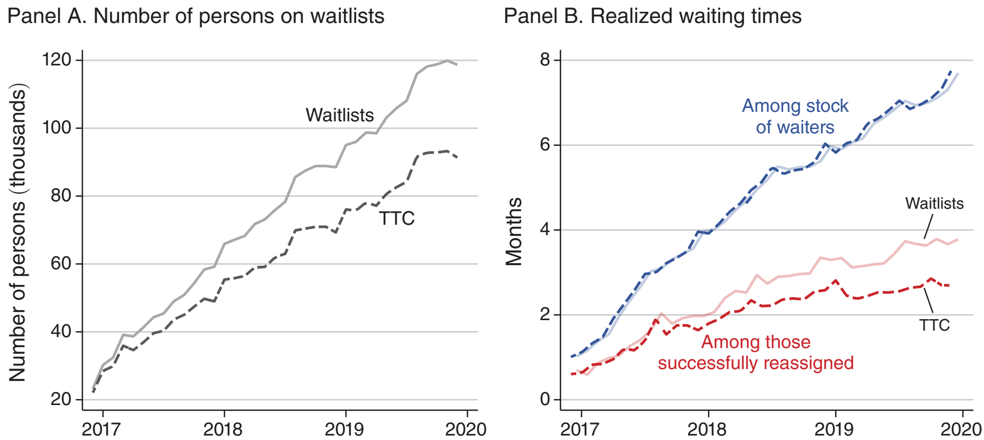

Figure 2. Results of Running TTC on Historical Data

Notes: The figure shows outcomes from a mechanical simulation in which TTC is run each month 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 stock of patients standing on waitlists each month (blue) and the average waiting time among the flow of patients who were successfully reassigned from a waitlist each month (red).

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 through a simple mechanical simulation of 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 as observed in the data. We then compare the number of reassignments and waiting times to those under the status quo mechanism. Supplemental Appendix B.1 provides additional details about the implementation of this analysis.

Figure 2 presents the results of this exercise, comparing the status quo waitlists mechanism (Waitlists) to the counterfactual monthly implementation of 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 shows the average waiting time among the flow of individuals successfully reassigned from 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, it does not offer a Pareto improvement relative to the status quo. Some patients 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. Supplemental Appendix B.3 provides stylized examples of this phenomenon. Supplemental Appendix Figure E.4 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 deprioritized relative to the status quo.

While this analysis suggests that there could be significant welfare gains from introducing TTC, it has two important limitations. First, the simulation holds patient behavior fixed. If the implementation of TTC changed the distribution of waitlist lengths, patients might have requested different GPs. Indeed, analysis in Supplemental 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. A second limitation 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. 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. Before moving to these more realistic simulations, we first estimate the primitives on which they will be based.

II. Model of Demand for GP Switching

Section IIA presents a model of patient demand to switch GPs. 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 to switch to. Section IIB then presents a belief model under which patients map waitlist lengths into beliefs about waiting time, which is an important input into patients' choice over GPs.

A. *Attention*, *Preferences*, *and Choice*

Patients are indexed by i, GPs by j, and time by t. Time is continuous. At any given time t, patient i has observable attributes X_{it}, and her preferences over GPs are represented by the indirect flow utility she would derive from being assigned to each GP, denoted $\mathbf{v}_{it} \equiv \left(v_{i1t}, \ldots, v_{iJt}\right) \in \mathbb{R}^J$. Across time, she discounts future flow utilities at exponential rate ρ.

As indicated in Figure 1, patient requests to switch GP are in fact quite rare, occurring for each patient on average once every 13 years. We therefore need to take a stand on how to interpret this infrequency. This pattern could simply indicate that patients for the most part have their most preferred GPs. Alternatively, patients could be aware that a preferred GP is available, but choose not to switch because switching is costly. Yet alternatively, patients could be unaware that a preferred GP is available because they are not "paying attention." For our purposes, the important distinction between these explanations is that they have different implications for how switching behavior would respond to alternative mechanisms. Empirical relationships in our data, discussed in more detail below, lead us to prefer the third explanation. We therefore proceed within a model of exogenous "inattention." Patients actively consider switching GPs ("pay attention") only periodically, according to a stochastic and exogenous process. When attentive, patients make an endogenous, fully informed choice of GP.

We model the rate at which patients pay attention as following a memoryless Poisson process.[28] 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. The rate p_{it}^{λ} at which patients pay attention is thus a function of patient observables X_{it}. When a patient pays attention, two things happen: (i) She draws new preferences $\mathbf{v}_{it} \sim F(\cdot \mid X_{it})$ and (ii) she considers switching GPs.[29]

When considering switching, an attentive patient makes a discrete choice among GPs, including her own current GP. To make this choice, she evaluates the net present value of requesting to switch to a given GP, imagining that she will remain with that GP forever. Choosing a GP with available slots would allow the patient to switch GPs and realize these benefits immediately. Choosing a GP with a waitlist would instead delay the benefits to some point in the future.[30] If at time t, patient i is currently assigned to GP j_0 and must wait $T \geq 0$ periods to be reassigned to GP j, the net present value of requesting GP j is

$$NPV_{ijt} = \int_{\tau=t}^{t+T} e^{-\rho(\tau-t)} v_{ij_0t} d\tau + \int_{\tau=t+T}^{\infty} e^{-\rho(\tau-t)} v_{ijt} d\tau \tag{1}$$
$$= \frac{1}{\rho}\Big[v_{ij_0t} + e^{-\rho T}\big(v_{ijt} - v_{ij_0t}\big)\Big].$$

The first line in equation (1) shows that the net present value of requesting a given GP can be decomposed into two parts. The first is the (dis)utility derived from staying with the current GP j_0 between now and time T, while standing on the waitlist for the new GP j. The second is the utility derived from being assigned to the new GP j from time T and forever after. The second line of equation (1) shows how this expression can be simplified. The patient is essentially maximizing the incremental utility derived from GP j relative to GP j_0, discounted T periods into the future.

The patient formulates beliefs, discussed in Section IIB below, about waiting times T_{ij} for each GP. Given these beliefs, and letting $\mathcal{J}$ denote the set of all GPs, the attentive patient's choice problem can now be written

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

[28] Because our data are at the monthly level, we will estimate monthly attention probabilities and assume that patients can only be attentive once within a month. In reality, patients do make GP selections continuously throughout the month, and our counterfactuals will simulate within-month patient arrival times.

[29] This formulation treats a patient's flow payoff from their GP as constant between attention spells. Of course, a patient's assigned GP only matters in an ex post sense at moments when they wish to consume health care. The specification of flow payoffs here abstracts away from specific health episodes and GP visits and should be interpreted in an ex ante sense. When choosing a GP, a patient does not necessarily know when they will want to see their doctor, but when they do, it will be valuable to be able to be assigned to a doctor who is convenient and pleasant to visit, and who provides good care.

[30] If a patient chooses her own current GP, she does not need to wait on the waitlist and can enjoy that assignment immediately.

This choice problem maximizes the value of being assigned to GP j after some waiting time T_{ij}, accounting for uncertainty in waiting time 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 discounting of the future benefit from switching. An attentive patient will therefore be predicted to *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 times. Our model thus interprets any switch request in the data as implying both that the consumer received an attention shock and 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 their 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 less preferred 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 perceived waiting times 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 infrequent within patient. This choice is not without loss of generality. A model in which there is a cost of making a switch request would predict that more patients would request to switch should aggregate waiting times fall, while a model of exogenous inattention would not.[31] 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 in the reassignment algorithm made during our sample period. Supplemental Appendix B.5.1 presents this analysis. We do not find any evidence that patients are more likely to request to switch when waitlists for nearby GPs 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.[32]

[31] 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).

[32] 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.

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 and because the identity of a patient's current GP does not affect their ability to switch GPs under the current system. 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 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 a few reasons. Since switch requests are rare, 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. Furthermore, patients often fail to fully exploit strategic opportunities even within the 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, they are rare—simultaneous open switches and waitlist joins occur in only 3 percent of patient-months in which a switch request was made. Finally, we test whether patients are more likely to request to switch GP in regions where the waitlists are growing more quickly (Supplemental Appendix B.5.2). We do not find evidence of additional switching when the value of future switching opportunities is falling more quickly.

Finally, we ignore the value of a long-term relationship with your GP. In 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 the mechanism design changes we study would have limited impact on patient-GP relationship capital.

B. *Waiting Time Beliefs*

It remains to specify patients' information and beliefs about waiting time. 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 propose a tractable model of beliefs that approximates the structure of a first-come, first-served queue in a way that depends on a small number of parameters that 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 w, a patient's total waiting time T can be thought of as the sum of their waiting time in each position on the waitlist until they are reassigned. The patient must first move from position w to position $w-1$, which takes time T_w; then from $w-1$ to $w-2$, which takes time T_{w-1}; and so on up to the time it takes to move from position 1 to being assigned, which takes time T_1. So, a patient can form beliefs about their total waiting time by forecasting their waiting time in each position and then summing across positions. The time the patient spends in position s 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 ahead on the waitlist abandon it before being reassigned. Incumbent enrollees may depart their panel in the event of death, emigration, or a switch to another GP. Patients standing on waitlists may abandon it in the event of death, emigration, or actively removing themselves from the waitlist.

We assume that patients perceive that slot vacancies and waitlist abandonments follow independent Poisson processes in continuous time. That is, patients believe that each slot on a GP's panel becomes vacant at exponential rate η, and each waiting patient abandons the waitlist they are on at exponential rate κ.[33] If these processes are independent, then the time T_{js} a patient spends in position s waiting for GP j (before moving to the next position, $s-1$) follows an exponential distribution with parameter $N_j\eta + (s-1)\kappa$, where N_j is GP j's panel cap. A patient's expected total waiting time when entering GP j's waitlist at position w is

$$T_j \equiv \sum_{s=1}^{w} T_{js} \qquad T_{js} \overset{iid}{\sim} \exp\big(N_j\eta + (s-1)\kappa\big).$$

For a given patient, T_s and $T_{s'}$ are independent for $s \neq s'$. Therefore, a patient's total expected waiting time is

$$\mathbb{E}\big[T_j \big| w\big] = \mathbb{E}\left[\sum_{s=1}^{w} T_{js}\right] = \sum_{s=1}^{w} \frac{1}{N_j\eta + (s-1)\kappa}, \tag{3}$$

and the expected discount factor is

$$\mathbb{E}\big[e^{-\rho T_j} \big| w\big] = \prod_{s=1}^{w} \mathbb{E}\big[e^{-\rho T_{js}}\big] = \prod_{s=1}^{w} \frac{N_j\eta + (s-1)\kappa}{\rho + N_j\eta + (s-1)\kappa}, \tag{4}$$

where equations (3) and (4) are derived using the facts that the waiting times in each position T_{js} are independent and exponentially distributed.

This belief model embeds three 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.[34] 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

[33] Of course, these rates are an equilibrium outcome of the reassignment mechanism used in the system. Though we condition on them in estimation, we will endogenize patients' beliefs about these rates in counterfactuals.

[34] In addition, at least ten slots must be available before any patients are assigned from the waiting list. This is done to provide a buffer in case there are multiple births to patients already on the GP's panel.

move. Third, patients believe that the parameters η and κ are constant across GPs. In principle, these parameters could vary across GPs with calendar time or with other factors. Given that in practice patients have limited information about the mapping from waitlist lengths to waiting times and are likely unaware of how this differs for individual GPs, we believe our belief model is a reasonable approximation.

III. Estimation

A. *Sample and Parameterization*

Geographic Subsample.—For computational tractability, we estimate demand 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.[35] Its population accounts for about 8 percent of the country. In terms of time period, we focus on the 12-month period from December 2018 to November 2019, by which time GP waitlists were well established. Supplemental Appendix Table E.2 provides a comparison of key demographic characteristics 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 restrict patient choice sets. Our baseline choice set definition is a driving time radius of 60 minutes around a patient's municipality of residence.[36] Details about the construction of the demand estimation sample are provided in Supplemental Appendix A.2. Our final demand estimation sample represents 4,213,049 patient-months (379,330 unique patients) and 457 unique GP panels.

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

$$\lambda_{it} \overset{iid}{\sim} \text{Bernoulli}\Big(p^{\lambda}(X_{it})\Big); \tag{5}$$

$$v_{ijt} = -d_{ijt} + \delta_j + X_{ijt}\beta + \epsilon_{ijt}. \tag{6}$$

The attention shock λ_{it} is drawn independently each period according to a probability $p^{\lambda}(X_{it})$ that depends on patient observables. Motivated by regression evidence in Supplemental Appendix B.4, X_{it} includes patient demographics (age, gender, permanent residency status, health status), whether the patient has recently or will imminently move, and if so, the distance of the move.[37] An attentive patient $(\lambda_{it} = 1)$

[35] Of the 11 administrative regions (*fylke*) in Norway, Trondelag has the lowest outside-region GP enrollment. Only 1.7 percent of residents of Trondelag enroll with a GP outside of the region, compared to 7.7 percent of residents of the Oslo region. Supplemental Appendix Figure E.5 provides a map of Norway and of Trondelag.

[36] 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.

[37] Other patient demographics, including income and education, are not included in the attention or choice models. We found in exploratory analyses that these variables had limited explanatory power for switching frequency or GP choice conditional on the other variables included in our model.

draws new taste shocks $\epsilon_{ijt} \overset{iid}{\sim} N\left(0, \sigma_{\epsilon}^{2}\left(X_{ijt}\right)\right)$ for each GP and then makes a GP choice. Inattentive patients simply remain with their current GP and retain their current taste shocks. The flow payoff v_{ijt} from being assigned GP j depends on 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. Motivated by evidence in Supplemental Appendix B.2, X_{ijt} includes indicators for whether the patient and GP are of the same gender and same age, and we allow the degree of age/gender homophily to vary across patient demographics and health status. We allow the variance of the idiosyncratic taste shock and GP fixed effects to vary by patient age and residency status.[38]

When reporting results, we normalize scale in the model by fixing patients' (dis)taste for travel time at -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). We normalize location by setting the fixed effect for one GP to zero. This leaves us with the following model parameters to estimate: $\left\{\rho, \delta, \beta, \sigma_{\epsilon}, p^{\lambda}\right\}$.

B. *Estimation Procedure and Identification*

We estimate the model in two steps. We first estimate the parameters governing patient beliefs using the empirical analogs of the objects described in Section IIB. We then jointly recover the attention and preference parameters that best describe the observed data given our model and patient beliefs.

Waiting Time Beliefs.—We estimate the belief parameters $\left(\eta, \kappa\right)$ using all waitlist-month observations in 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 the following belief parameters, which should be interpreted as monthly rates:

$$\text{panel slot vacancy rate:} \qquad \eta = 0.00181;$$

$$\text{waitlist abandonment rate:} \qquad \kappa = 0.01704.$$

The average waitlist length over these months was 28, and the average panel cap was 1,084. Using our formula for expected waiting time (equation (3)), 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 waiting time

[38] In estimation, we normalize the standard deviation of the idiosyncratic taste shock to 1, allow the coefficient on distaste for travel time to vary by patient age and residency status, and estimate common GP fixed effect coefficients. This specification implies heterogeneous variances of idiosyncratic tastes and GP fixed effects when flow payoffs are normalized by patients' distaste for travel time.

among patients on waitlists as of December 2019 was 7.9 months, which is a lower bound on the waiting time 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 rate 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) $$\delta_b, \beta_b \,|\, \mathbf{v}_{b-1}, \sigma_{\epsilon,b-1}^2, \mu_0, \Sigma_0;$$

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

(iii) $$\rho_b \,|\, \delta_b, \beta_b, \sigma_{\epsilon,b}^2, y, Z;$$

(iv) $$\lambda_b \,|\, p_{b-1}^\lambda, \delta_b, \beta_b, \sigma_{\epsilon,b}^2, y, Z;$$

(v) $$p_b^\lambda \,|\, \lambda_b, \alpha, \varphi;$$

(vi) $$\mathbf{v}_b \,|\, \lambda_b, \delta_b, \beta_b, \sigma_{\epsilon,b}^2, \rho_b, y, Z.$$

Supplemental 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.[39] 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 (i) the distribution of flow payoffs, (ii) the discount rate, and (iii) the attention parameters. Beliefs are assumed known.

First, suppose the discount rate ρ is known and attention is observed. The distribution of flow payoffs is nonparametrically 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 health care, 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.

[39] 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.

Second, given the distribution of flow payoffs, the discount rate is identified by the sensitivity of patients' choices to waitlist lengths (Waldinger 2021).[40] 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 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. Our empirical specifications therefore estimate a fixed effect for each GP, effectively isolating sensitivity to waiting time conditional on GP quality. 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 first-come, first-served 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, it would lead us to underestimate the discount rate.

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, some 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.

C. *Estimates*

We estimate three specifications. All specifications allow attention probabilities to differ flexibly by patient age, gender, whether the patient has a chronic health condition, and whether the patient is about to or recently moved. In the first specification, flow payoffs depend only on distance and interactions between patient and GP age and gender, where the strength of age and gender homophily varies flexibly by patient chronic health status. The second specification adds GP fixed effects to allow for persistent differences in GP desirability. The third specification allows the standard deviations of the idiosyncratic taste shock and GP fixed effects to vary by patient age and permanent residency status.

Table 3 presents the parameter estimates. Column 1 reports considerable sensitivity to waiting time, with an annual discount factor of 0.942. When GP fixed effects

[40] Though we estimate a common discount rate for all patients, one could allow it to depend on 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.

Table 3—Preference Parameter Estimates

	(1)		(2)		(3)	
Variable	β	SE	β	SE	β	SE
Annual discount factor	0.942	0.002	0.916	0.009	0.915	0.009
Travel time (minutes)	−1.000[a]		−1.000[a]		−1.000[a]	
SD GP fixed effects			19.761	1.032		
SD GP fixed effects, temp. res.					19.084	1.382
SD GP fixed effects, perm. res. age $\leq$ 45					21.572	1.240
SD GP fixed effects, perm. res. age $>$ 45					18.199	1.234
Male GP						
× Temporary resident	−3.070	0.338	0.000[a]		0.000[a]	
× Perm. res. female, no chron., age $\leq$ 45	−1.122	0.367	−4.906	0.465	−5.368	0.518
× Perm. res. female, no chron., age $>$ 45	−0.484	0.410	−4.779	0.987	−4.534	0.942
× Perm. res. female, chronic, age $\leq$ 45	−1.902	1.322	−6.244	1.456	−6.649	1.403
× Perm. res. female, chronic, age $>$ 45	0.325	0.426	−3.924	0.562	−3.616	0.561
× Perm. res. male, no chron., age $\leq$ 45	3.799	1.299	−0.104	1.226	−0.315	1.356
× Perm. res. male, no chron., age $>$ 45	5.075	1.444	0.914	1.338	0.641	1.174
× Perm. res. male, chronic, age $\leq$ 45	6.182	0.576	0.436	0.871	0.528	0.920
× Perm. res. male, chronic, age $>$ 45	6.136	0.518	0.397	0.840	0.212	0.814
GP age $>$ 45						
× Temporary resident	−1.548	0.359	0.000[a]		0.000[a]	
× Perm. res. female, no chron., age $\leq$ 45	−0.784	0.387	−2.551	0.366	−2.754	0.449
× Perm. res. female, no chron., age $>$ 45	−0.012	0.433	−2.211	1.000	−1.990	0.852
× Perm. res. female, chronic, age $\leq$ 45	−2.138	1.356	−4.345	1.359	−4.567	1.588
× Perm. res. female, chronic, age $>$ 45	−0.168	0.455	−2.230	0.417	−2.067	0.483
× Perm. res. male, no chron., age $\leq$ 45	−2.028	1.321	−3.297	1.089	−3.676	1.128
× Perm. res. male, no chron., age $>$ 45	−0.380	1.397	−1.779	1.269	−1.661	1.066
× Perm. res. male, chronic, age $\leq$ 45	−0.977	0.555	−3.129	0.755	−3.379	0.827
× Perm. res. male, chronic, age $>$ 45	−0.223	0.534	−2.219	0.671	−2.064	0.648
SD idiosyncratic shock	14.194	0.263	11.698	0.382		
SD idiosyncratic shock, temp. res.					11.116	0.607
SD idiosyncratic shock, perm. res. age $\leq$ 45					12.567	0.494
SD idiosyncratic shock, perm. res. age $>$ 45					10.603	0.574

Notes: The table reports parameter estimates from the GP choice model described in Section II. 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 and GP fixed effects to differ by residency status and age.

[a] By normalization.

are added in columns 2 and 3, this value falls to 0.915, reflecting the fact that more desirable GPs have longer waitlists. We estimate strong homophily by gender and age. By gender, estimates in column 1 imply that a female patient under age 45 without a chronic condition would travel 4.9 minutes farther than a male patient under 45 to see a female GP (5.6 minutes for female versus male patients over 45). Estimates imply slightly greater gender homophily for patients with a chronic condition.[41] By age, patients under 45 would travel approximately one to two minutes farther than a

[41] In exploratory analyses, we did not find that patients' GP choices depended on waiting time or distance in ways that differed strongly by health status. We therefore included health status interactions with demographic preferences but not in other parts of the preference model. One possible interpretation of these patterns is that for patients with a chronic health condition, specialist physicians play a relatively more important role than GPs in health care delivery.

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 yield similar magnitudes.

Column 3 allows the variance of the idiosyncratic taste shock and GP fixed effects to vary by patient age and residency status, which can be equivalently thought of as allowing for heterogeneity in the distaste for travel time relative to GP quality and idiosyncratic factors. We find that the distaste for travel time is highest for older permanent residents, and lowest for younger permanent residents. In both columns 2 and 3, we estimate 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 19.8 minutes, and a standard deviation of idiosyncratic shock of 11.7 minutes. In column 3, the standard deviation of GP fixed effects is between 18.2 and 21.6 minutes, and the standard deviation of idiosyncratic shock is between 10.6 and 12.6 minutes.[42]

Table 4 presents parameter estimates for the attention model. Estimates are similar across specifications, so we focus on column 3. Our estimates imply that nonmovers consider switching GPs far less often than patients who have recently moved, consistent with observed switching patterns (see Table 1). Among nonmovers, temporary residents pay attention most often—1.17 percent chance per month (approximately once every 6.5 years)—while older men consider switching just once every 25 years. Conditional on age and gender, patients with a chronic health condition make active choices 20–50 percent more frequently than patients without a chronic condition.

Patients who have moved consider switching an order of magnitude more often. We estimate separate attention probabilities for short and long moves, where short means less than 30 minutes' drive time between origin 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 six months), reflecting the fact that we see a large immediate impact of moving on the probability of switching GPs in the data (see Supplemental 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 18.33 percent chance of considering switching GPs. The probability remains high, but falls to 6.85 percent per month in the six months following the move. Permanent residents exhibit a qualitatively similar pattern. 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 eight months around that move. If the move were over 30 minutes, this rises to 38 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.

[42] 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 4—MONTHLY ATTENTION PROBABILITY ESTIMATES

	(1)		(2)		(3)	
Variable	p^{λ}	SE	p^{λ}	SE	p^{λ}	SE
No recent or imminent move						
× Temporary resident	1.18	0.02	1.18	0.03	1.17	0.03
× Perm. res. female, no chron., age $\leq$ 45	0.72	0.01	0.72	0.02	0.72	0.02
× Perm. res. female, no chron., age $>$ 45	0.41	0.01	0.46	0.02	0.47	0.02
× Perm. res. female, chronic, age $\leq$ 45	0.97	0.02	1.08	0.03	1.08	0.04
× Perm. res. female, chronic, age $>$ 45	0.51	0.01	0.57	0.01	0.57	0.01
× Perm. res. male, no chron., age $\leq$ 45	0.44	0.01	0.44	0.01	0.44	0.01
× Perm. res. male, no chron., age $>$ 45	0.30	0.01	0.30	0.01	0.30	0.01
× Perm. res. male, chronic, age $\leq$ 45	0.48	0.02	0.60	0.08	0.60	0.08
× Perm. res. male, chronic, age $>$ 45	0.33	0.01	0.39	0.03	0.40	0.03
Moved $\leq$ 30 minutes, this or next month						
× Temporary resident	6.90	1.14	6.90	1.14	6.93	1.14
× Perm. res. female	4.51	0.39	4.51	0.38	4.51	0.38
× Perm. res. male	3.32	0.33	3.32	0.33	3.32	0.34
Moved $\leq$ 30 minutes, prev. 6 months						
× Temporary resident	3.10	0.54	3.09	0.54	3.08	0.53
× Perm. res. female	2.12	0.17	2.11	0.17	2.11	0.17
× Perm. res. male	1.58	0.15	1.58	0.14	1.58	0.14
Moved $>$ 30 minutes, this or next month						
× Temporary resident	18.25	2.24	18.30	2.26	18.33	2.29
× Perm. res. female	8.62	0.64	8.59	0.66	8.62	0.65
× Perm. res. male	37.81	1.13	29.67	13.73	29.89	13.61
Moved $>$ 30 minutes, prev. 6 months						
× Temporary resident	6.81	1.08	6.87	1.11	6.85	1.11
× Perm. res. female	3.42	0.27	3.42	0.28	3.42	0.28
× Perm. res. male	2.60	0.24	2.61	0.23	2.61	0.23

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 change their municipality of residence in the six months prior to or one month after the current month. After a patient has been attentive during a given move spell, they are treated as a nonmover during the rest of that move spell.

IV. Counterfactual Simulations and Welfare

We now use our estimates to predict equilibrium outcomes under alternative waitlists mechanisms. Section IVA describes the simulated dynamic economy and defines our counterfactual mechanisms. Section IVB defines an equilibrium. Section IVC presents equilibrium outcomes, and Section IVD discusses extensions.

A. *Counterfactual Simulations*

Dynamic Economy.—We consider an economy with a finite set of patients and GPs. The set of GPs is given by $J = \{j_1, \ldots, j_m\}$. Each GP's type includes their age, gender, office location, panel cap, and fixed effect δ_j. All GP characteristics are fixed over time. The set of patients is given by $I = \{i_1, \ldots, i_n\}$. Each patient's type $x \in \mathcal{X}$ includes their age, gender, residency status, chronic health status, and location of residence (one of the 45 municipalities in the Trondelag region). Patient characteristics evolve over time according to a stationary Markov process

$M: \mathcal{X} \rightarrow \mathcal{X}$. Patients also have a chance of dying, in which case they immediately experience a "rebirth" 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.[43]

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

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

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

(ii) (*Attention*) Patients draw attention shocks according to the probabilities $p^{\lambda}(x)$ reported in Table 4.

(iii) (*GP Choice*) Attentive patients sequentially arrive to the mechanism. Upon entry, they draw new preferences (ϵ_{ij}) for all GPs, formulate flow indirect utilities v_{ij} based on parameter estimates in Table 3, and report their GP request(s) to the mechanism. Patients who request to stay with their current GP or switch to an open GP are immediately reassigned and exit the mechanism. Patients who request a full GP are placed on the waitlist.

(iv) (*Matching Algorithm*) The patient-GP matching algorithm is executed. Patients that are successfully reassigned exit the mechanism. All other patients 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. Supplemental Appendix Table E.3 summarizes these realizations.

Allocation Mechanisms.—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 algorithm, just as in the mechanical simulations from Supplemental Appendix 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 is described below and formally defined in Supplemental Appendix Section D.2.

[43] 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.

Waitlists: This is the mechanism currently used in Norway. The matching algorithm iteratively advances waitlists to fill open GP slots until there are no patients waiting for GPs with open slots. Patients are prioritized in order of waiting time.

Waitlists with TTC: This mechanism is the same as Waitlists but with a different matching algorithm. This matching algorithm consists of first running the Waitlists matching algorithm and then running the TTC algorithm on any remaining patients. Patients are prioritized in order of waiting time.

Waitlists with TTCP: This mechanism is identical to TTC except that the matching algorithm prioritizes patients with undersubscribed GPs above patients with oversubscribed GPs. Within each group, patients are still prioritized in order of waiting time. A GP is undersubscribed if it has at least one open slot on its panel.[44] TTCP attempts to address some of the adverse distributional consequences from introducing TTC. Rather than preventing waiting patients from exchanging their GPs, TTCP gives patients who are unlikely to benefit from cycles priority for other open GP slots.

Waitlists with DA: This mechanism is the same as Waitlists but with a different matching algorithm, namely the patient-proposing DA algorithm. 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 waiting time priorities. No patient may "jump" to the front of the queue and be reassigned to a GP for whom another patient was waiting longer.[45] Nevertheless, by strictly respecting waiting time priority, DA may substantially limit trading opportunities relative to TTC.

Measures of Welfare.—We consider two cardinal measures of patient welfare. The first is simply the mean (or median) per-period flow payoff v_{ijt} experienced by each patient in the economy, weighing all patients and all time periods equally. The mean flow payoff would be relevant to a utilitarian planner interested in maximizing the present discounted value of future payoffs in the economy.

Our second measure isolates how the value of switching opportunities changes across mechanisms. It is based on patients' perceived net present value of making an optimal GP choice. Supplemental Appendix D.4 provides details on how this measure is constructed.[46] We scale the net present value by patients' discount rate ρ

[44] We classify a patient's current GP as over-/undersubscribed at the moment they enter the mechanism. 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 modeling patient beliefs about this possibility would be highly complex.

[45] 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 first-come, first-served priority (i.e., generating envy).

[46] A complication with constructing such a measure in a dynamic economy is that the value of a patient's choice problem depends on their current GP, and 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 measure the value of an attentive patient's choice problem *holding the patient's current GP fixed* at its value under Waitlists.

(estimated to be 0.0075) to yield the equivalent perpetuity per-period flow payoff, as given by equation (7) below. This allows for comparison to the first welfare measure, which is also expressed in per-period payoffs. Aggregate welfare is then represented by the mean (or median) value across all attentive patient-months. Since this measure places more weight on time periods shortly after a patient receives an attention shock, it captures the idea that one might care more about improving patients' experience at the moment they wish to change GPs, rather than in all time periods equally. This measure also allows us to compare patient welfare as a function of characteristics that are endogenous to the mechanism, including whether the patient currently has an over- or undersubscribed GP.

B. *Beliefs*, *Decisions*, *and Equilibrium*

Let $\mathbf{w}_{it} = \{w_{i1t}, \ldots, w_{iJt}\} \in \mathbb{N}_+^J$ denote patient i's waitlist position if she were to select each GP at time t.[47] In all of our counterfactuals, attentive patients choose a GP by solving the choice problem in equation (2), conditioning their beliefs about waiting time on observed waitlist lengths:

$$\max_{j \in \mathcal{J}_{it}} \mathbb{E}\left[e^{-\rho T_{ij}} \middle| \mathbf{w}_{it}\right]\left(v_{ijt} - v_{ij_0t}\right). \tag{7}$$

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—that is, the expected discount factor function $\mathbb{E}\left[e^{-\rho T_{ij}} \middle| \mathbf{w}_{it}\right]$. 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 IIB.

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.[48] 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.

[47] If a GP has open slots or patient i is already on that panel, then i's waitlist position is zero. If the patient is already sitting on GP j's waitlist, she would retain that position.

[48] 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.

A patient's expected discount factor when entering GP j's waitlist at position s can now be written

$$\begin{aligned}(8)\quad \mathbb{E}\left[e^{-\rho T_{ij}}\middle| j_0, s\right] &= \frac{m_{js}+\chi_{j_0 s}}{\rho+m_{js}+\chi_{j_0 s}}\Bigg(\frac{\chi_{j_0 s}}{m_{js}+\chi_{j_0 s}} \\ &\qquad + \frac{m_{js}}{m_{js}+\chi_{j_0 s}}\,\mathbb{E}\left[e^{-\rho T_{ij}}\middle| j_0, s-1\right]\Bigg) \\ &= \frac{\chi_{j_0 s}}{\rho+m_{js}+\chi_{j_0 s}} + \frac{m_{js}}{\rho+m_{js}+\chi_{j_0 s}}\,\mathbb{E}\left[e^{-\rho T_{ij}}\middle| j_0, s-1\right],\end{aligned}$$

where $m_{js} \equiv N_j\eta + (s-1)\kappa$ is the rate at which the patient moves forward in the queue.[49] Equation (8) 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: $\chi_{j_0 s} = \mathbf{1}[j_0 \text{ oversub.}]\exp\big(\chi_0 + \chi_1 \log(s/N)\big)$.[50] In addition to being parsimonious, Supplemental Appendix Figure E.6 shows that this parameterization does a good job fitting the probabilities of being reassigned through a cycle in our equilibrium simulation of TTC. 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. Supplemental Appendix D.3 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. Supplemental Appendix Table D.2 reports equilibrium beliefs.

C. *Equilibrium Outcomes*

Table 5 reports equilibrium outcomes from a five-year window at the end of our simulation period, when the economy has reached a stationary equilibrium. In the long-run stationary equilibrium of Norway's current mechanism (Waitlists), 11 percent of patients are on a waitlist, and 81.3 percent of GPs have a waitlist. Patients' expected waiting time to switch to the average GP (including zeros for GPs without waitlists) would be 18.8 months. Each month, an average of 2,368 patients receive attention shocks and consider switching GPs (see Supplemental Appendix

[49] 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 $\mathbb{E}\left[e^{-\rho T_{ij}}\middle| j_0, s-1\right]$.

[50] In the TTCP mechanism, we additionally 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.

Table 5—Outcomes under Alternative Mechanisms

	Waitlists	TTC	DA	TTCP
GP waitlists				
Percent of population on a waitlist	11.0	10.4	11.0	11.0
Percent of GPs with a waitlist	81.3	78.9	81.3	79.9
Mean $\mathbb{E}$(wait time) | curr. GP undersub.	18.8	23.9	18.7	19.1
| curr. GP oversub.	18.8	12.0	18.7	13.5
Attentive patient choices				
Mean $\mathbb{E}$(wait time) at chosen GP	18.8	15.6	18.8	16.6
Percent waitlist joins	83.1	83.9	83.1	84.1
| curr. GP undersub.	79.1	75.3	79.0	77.6
| curr. GP oversub.	83.8	85.5	83.8	85.1
True pref. rank of chosen GP	1.80	1.63	1.80	1.63
| curr. GP undersub.	1.9	2.2	1.9	2.0
| curr. GP oversub.	1.8	1.5	1.8	1.6
Realized assignments				
Travel time to current GP, mean (med.)	14.8 (6.4)	14.6 (6.4)	14.8 (6.4)	14.8 (6.4)
Percent with same gender GP | young female	58.1	58.9	58.1	59.1
| young male	60.0	59.2	60.0	59.0
Welfare				
Flow payoff from current GP, mean (med.)	—[a]	0.72 (0.42)	0.01 (−0.00)	0.40 (0.26)
Perpetuity equiv. of GP choice, mean (med.)	—[a]	1.54 (0.71)	0.01 (0.00)	1.36 (0.43)

Notes: The table reports statistics on outcomes generated in months 492–551 of the simulation, out of 600 total months. Statistics are first computed within month and then averaged across simulation months. $\mathbb{E}$(wait time) 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.

[a] By normalization.

Table E.3). Among these patients, 83.1 percent choose to join 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 18.8 months. 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 1.80.

Our model predicts significantly longer equilibrium waitlists and waiting times than those observed in our data, reflecting the fact that the waitlists were still growing rapidly at the end of our sample period. As of August 2025, the percent of the population standing on a GP waitlist seemed to have stabilized at 6.5 percent, more than twice the number at the end of our sample period.[51] Our model's prediction that the waitlists would stabilize at a much higher level is therefore at least qualitatively consistent with what actually occurred.[52]

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 15.6 months, and a smaller share of patients are on a waitlist at a given time. Measured by realized flow payoffs, mean patient welfare increases by the

[51] The number of patients currently standing on each GP's waitlists is publicly available at https://tjenester.helsenorge.no/bytte-fastlege.

[52] Of course, our counterfactual simulations hold fixed many aspects of Norway's health care system that changed in reality, including the COVID-19 pandemic. Our simulations also apply only to the Trondelag region. For these reasons, we would not expect our prediction of the share of patients on waitlists to perfectly align with the share observed in all of Norway.

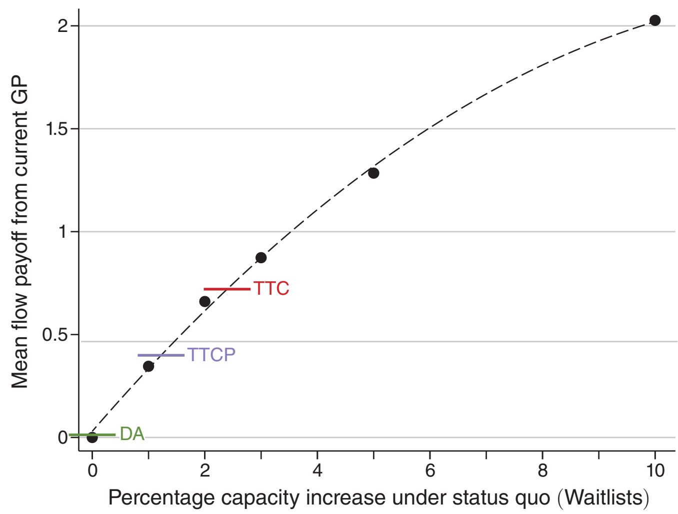

Figure 3. Comparison to Welfare Effect of Capacity Increases under Status Quo

Notes: The figure shows the welfare effect (in terms of mean experienced flow payoffs) of increasing capacity under the status quo mechanism (Waitlists). Panel capacity is increased by a fixed percentage for all GPs in the economy. We run the simulation for 1, 2, 3, 5, and 10 percent capacity increases (0 percent corresponds to the original Waitlists simulation). The dotted line shows a quadratic fit. For comparison, welfare associated with TTC, TTCP, and DA are also shown. In the limit, welfare implied by infinite GP capacity yields a mean flow payoff of 6.89.

equivalent of 0.72 minutes' driving time. More than one-quarter of these gains are attributable directly to patients being matched with closer GPs, with the remainder driven by improved match quality on both observable and unobservable dimensions. The equivalent perpetuity payoff perceived by the average attentive patient is 1.54 minutes.[53] These average improvements are reflected in patients' behavior: They become more likely to request their first-choice GP, with the average preference rank of chosen GPs falling to 1.63.

To aid in interpreting the magnitude of these welfare effects, Figure 3 plots the mean flow payoff that our model predicts would arise if GP panel caps were raised under the status quo mechanism (Waitlists). We find that the welfare gains from TTC are equivalent to a capacity increase of approximately 2.5 percent, suggesting that the impacts of simple market design changes can have benefits that are comparable to substantial investments in system capacity. The figure also shows that, unsurprisingly, additional capacity has diminishing returns. In the limit, welfare implied by infinite GP capacity yields a mean flow payoff of 6.89.[54]

[53] The difference between the experienced flow payoff and perceived perpetuity equivalent welfare measures reflects three things. First, the perceived perpetuity equivalent is the result of a hypothetical exercise in which each attentive patient is imagined to have the same *current* GP as they were assigned under Waitlists at any given time, even though their actual GP may be different under an alternative mechanism. Second, the perceived perpetuity equivalent reweights patient months relative to the experienced flow payoff measure because it is calculated in *attentive* patient months only, and from there discounts future payoffs. Since attentive patients are weakly improving their situation (relative to being inattentive), welfare gains tend to be larger according to the perpetuity equivalent measure than according to the flow payoff measure. Finally, patients are overoptimistic 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.

[54] In reality, GP quality would likely be affected if panel caps were eliminated. Infinite capacity is presented only as a benchmark of the maximal possible welfare that could be implied by our model.

Table 6—Distribution of Welfare Gains Relative to Waitlists

	Frac. of attn. pats.	TTC	DA	TTCP
All attentive patient-months	1.00	1.5 (0.7)	0.0 (0.0)	1.4 (0.4)
Patient demographics				
Perm res. male $<$ 45	0.20	1.8 (0.8)	0.0 (0.0)	1.6 (0.5)
Perm res. male $\geq$ 45	0.16	1.4 (0.6)	0.0 (0)	1.2 (0.3)
Perm. res. female $<$ 45	0.28	1.7 (0.9)	0.0 (0.0)	1.5 (0.6)
Perm. res. female $\geq$ 45	0.22	1.3 (0.7)	0.0 (0.0)	1.1 (0.4)
Temporary resident	0.14	1.5 (0.6)	0.0 (0)	1.4 (0.3)
Months since move				
Moved in last year	0.16	3.0 (1.0)	0.0 (0)	2.8 (0.6)
Moved over a year ago	0.35	1.4 (0.7)	0.0 (0.0)	1.3 (0.4)
Never moved	0.49	1.1 (0.7)	0.0 (0.0)	0.9 (0.4)
Geographic location				
Rural	0.19	3.5 (0.8)	0.0 (0)	3.3 (0.6)
Suburban	0.37	1.2 (0.7)	0.0 (0)	1.1 (0.5)
Urban (Trondheim)	0.45	1.0 (0.7)	0.0 (0.0)	0.8 (0.4)
Current GP oversubscribed				
No	0.16	−0.6 (−0.3)	0.0 (0)	0.1 (0)
Yes	0.84	1.9 (0.9)	0.0 (0.0)	1.6 (0.5)
Chronic health condition				
No	0.55	1.6 (0.7)	0.0 (0.0)	1.4 (0.4)
Yes	0.31	1.5 (0.7)	0.0 (0.0)	1.3 (0.5)

Notes: This table reports mean (and median) welfare in terms of the perceived perpetuity equivalent of attentive patients among different patient subgroups. Patient type is defined at the moment the patient is attentive. Statistics are calculated across all attentive patient-months in months 492–551 of the simulation, holding each patient's current GP fixed at their assignment under Waitlists.

Distributional Implications.—Table 6 summarizes welfare outcomes using the perceived perpetuity equivalent by subgroups of attentive patients. In terms of demographics, the welfare gains from TTC conditional on being attentive are concentrated among younger patients. Female patients also derive especially large benefits because they are more likely to be attentive than male patients. Patients who moved in the last 12 months also realize large benefits from TTC (3.0 minutes), reflecting the fact that these patients tend to be highly mismatched with their current GP and thus stand to gain substantially from shorter waiting times. However, patient moves are not the only dynamic driving patient-GP mismatch; patients who have never moved also benefit (1.1 minutes). In terms of geography, patients in both urban and rural areas benefit, but the gains are largest among rural patients (3.5 minutes), who tend to face the longest travel times and thus have the greatest potential for geographic mismatch. Interestingly, gains are similar for patients with and without a chronic health condition. Although chronic patients tend to consider switching more frequently than nonchronic patients of the same age and gender, the value of their GP choice increases by a similar amount. Finally, it is worth noting that for most groups, welfare gains are also positive and economically significant at the median, suggesting that gains are widely spread rather than concentrated among a minority of patients.

Although TTC benefits the majority of patients, it has particularly uneven consequences along an important dimension—whether a patient's current GP is oversubscribed. Table 5 shows that while TTC reduces waiting times for the average GP by

6.8 months (from 18.8 to 12.0) for patients with oversubscribed GPs, it *increases* waiting times (from 18.8 to 23.9 months) for patients with undersubscribed GPs. Patients' GP choices reflect the disparity along this dimension. Patients with undersubscribed GPs become less selective under TTC, choosing GPs with an average preference rank of 2.2 (instead of 1.9 under Waitlists), and joining waitlists less frequently. Patients with oversubscribed GPs, on the other hand, 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 in patient welfare. Table 6 reports that attentive patients with an undersubscribed GP are substantially worse off under TTC—the perpetuity equivalent of 0.6 minutes' travel time. Although these harms are offset by larger gains for patients with oversubscribed GPs, they may raise equity concerns. 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 first-come, first-served priority; patients may only swap GPs if they are both at the front of their respective waiting list. As a result, the waiting time reduction under DA is very small. This finding demonstrates a fundamental trade-off between eliminating envy—here, not allowing patients to "cut" in line—and finding additional gains from trade.[55]

TTCP instead attempts to directly advantage patients with undersubscribed GPs, without preventing feasible trades. Patients with undersubscribed GPs are prioritized 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 more than half of the welfare gains of TTC as measured by mean flow payoffs. Compared to TTC, more patients are standing on a waitlist at a given time, and patients expect to wait one month longer for their chosen GPs. For a patient with an oversubscribed GP, the expected waiting time for the average GP in their choice set rises from 12.0 to 13.5 months. However, this is offset by a much larger drop in expected waiting times for patients with undersubscribed GPs, from 23.9 to 19.1 months. These patients become more selective, requesting a GP with an average preference rank of 2.0 instead of 2.2, while patients with oversubscribed GPs become slightly less selective. The fourth column of Table 6 shows that patients with an undersubscribed GP are as well off under TTCP as under Waitlists, while patients with oversubscribed GPs are still better off by 1.6 minutes.

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, assignments to a vacancy under TTCP usually end with the assigned patient. Indeed, the rate at which the waitlist moves from the

[55] The result that DA heavily restricts gains from trade mirrors empirical findings in other studies, for example, Combe et al. (2022).

front is 30 percent lower under TTCP than under TTC (see Supplemental 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.

Despite yielding smaller improvements than TTC, the results from TTCP are encouraging insofar as they address some of the distributional concerns with introducing TTC in a market where some patients have more desirable endowments than others. The harms to patients with undersubscribed GPs are eliminated—they are slightly better off than under the status quo (see Table 6). At the same time, patients with oversubscribed GPs retain the benefits of being able to trade GPs, and are still considerably better off than under the status quo.

D. *Discussion and Extensions*

We find that simple changes to the assignment algorithm and priority rules of Norway's GP allocation mechanism can significantly reduce waiting times and increase patient welfare. In this section, we consider the implications of three extensions: not having waitlists, relaxing the limitation to a single waitlist, and relaxing incentive compatibility constraints. For all these cases, our ability to predict equilibrium outcomes is limited to some degree, so we view these results as suggestive only.

No Waitlists.—A natural question is whether introducing waitlists, as Norway did in 2016, in fact improves patient welfare. The answer is not obvious. For example, a patient who is highly mismatched with their current GP may prefer an environment in which more GPs are available immediately to one in which they have an expanded choice set but fewer immediately available options. Understanding the consequences of introducing waitlists is particularly relevant to the design of primary care systems because many systems that are otherwise similar to the one in Norway—including, to our knowledge, all HMOs in the United States—currently do not allow patients to join waitlists.

To attempt to evaluate this trade-off, we run a benchmark simulation, *No Waitlists*, in which an attentive patient may only choose among the GPs with open slots at the moment they consider switching. GPs are thus effectively rationed by "luck" due to stochastic availability, rather than through a first-come, first-served priority system. This exercise can be interpreted as a simulation of Norway's mechanism prior to 2016. However, we caution that outcomes arising from this simulation are not necessarily an equilibrium. In reality, patients would have an incentive to exert effort (e.g., frequently checking the website until a spot opens up) to obtain their desired GP. Nonetheless, this benchmark helps us assess whether Norway's introduction of waitlists in 2016 is likely to have significantly improved patient-GP match quality.

Supplemental Appendix Table D.3 summarizes outcomes under this simulation. We find that patients' mean flow payoff is just *above* its level under the status quo (0.10 minutes). Patients are reassigned to less-preferred GPs on average, but because waiting times are eliminated, patients spend less time mismatched with their current GP. This offsetting effect is especially valuable for patients with highly mismatched current GPs. It is therefore not so surprising that while mean welfare increases slightly, median welfare decreases by 1.43 minutes. While patients with a highly mismatched current GP benefit from the full elimination of waiting times, the majority of patients prefer an environment with more options and some waiting time. We interpret these results as suggesting that the desirability of introducing waitlists depends on how the gains for highly mismatched patients are weighed against the losses for patients who would prefer to wait for a more desirable option.

Joining Multiple Waitlists.—An important restriction in all of the mechanisms analyzed so far is that patients may only sit on one GP's waitlist at a time. This restriction limits the mechanism's ability to find gains from trade because patients may only express that one GP is preferred to their current assignment. Unfortunately, assessing the *equilibrium* implications of allowing patients to rank multiple GPs is challenging because such a change adds considerable computational complexity to a patient's choice problem.[56] We therefore instead consider a benchmark simulation, *Truthful TTC*, in which patients truthfully report their full preference list to the mechanism. This mechanism in effect allows patients to join an unlimited number of waitlists, but assumes they join the waitlists for *all* GPs preferred to their current GP. The mechanism is formally described in Supplemental Appendix D.5, and the results are reported in Supplemental Appendix Table D.3.

Qualitatively, Truthful TTC has similar impacts to No Waitlists: It nearly eliminates waiting times, but leads to lower postassignment match quality. Almost all patients can be reassigned to some preferred GP at the end of the same month in which they are attentive because most patients prefer many GPs to their current one (often including at least one undersubscribed GP), and because horizontal preference heterogeneity generates many feasible trades. The reduction in waiting times does come at the cost of a less preferred assignment. The average preference rank of a patient's chosen GP is 2.73 under Truthful TTC, compared to 1.80 under Waitlists and 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.

Quantitatively, however, the loss in match quality under Truthful TTC is smaller than under No Waitlists, and the resulting welfare gains greater, because Truthful TTC also finds gains from trade among waiting patients rather than only allowing assignments to vacancies. In fact, mean patient welfare is higher under Truthful TTC than under *any* of the focal mechanisms. In terms of realized flow payoffs, patients are on average 1.30 minutes better off relative to Waitlists (compared to 0.72 minutes under TTC). However, as under No Waitlists, these gains are concentrated among a minority of patients; the median patient is 0.61 minutes *worse* off. In contrast, TTC and TTCP modestly reduce waiting times while improving postreassignment match quality, yielding more widely distributed gains.

[56] In static implementations of DA or TTC with unrestricted preference lists, patients can do no better than truthfully reporting their ordinal preferences. In a dynamic implementation of these algorithms, "strategy-proofness" 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 that allow patients to rank a small number of GPs becomes 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.

Optimal Mechanisms.—Finally, a natural question is what an *optimal* allocation mechanism would look like in this environment. This question is challenging to answer. We are unaware of theoretical characterizations of optimal mechanisms in models with the types of agent and object heterogeneity considered here. Such mechanisms could involve complex choice and priority rules that depend on the current state of the entire GP system. In principle, it is still possible to solve computationally for an optimal mechanism, but this is an extremely high-dimensional problem in our model.

Despite these difficulties, we can make some progress toward assessing how large the gains from optimal mechanisms might be by solving a simpler problem. Specifically, we compute a *greedy first-best* allocation in which the planner can observe patients' preferences and computes the utilitarian-optimal reassignments among all patients who become attentive and wish to switch GPs each month. This allocation is not a true first-best because it ignores potential gains from having some patients *wait* for future arrivals of GP slots that are not available today. Nonetheless, this allocation helps us to assess the degree to which relaxing incentive compatibility and first-come, first-served priority constraints could improve the allocation.

Supplemental Appendix Table D.3 shows that the greedy first-best improves patient welfare by a mean flow payoff of 4.80. This amount, achieved at status quo capacity, is equal to a striking 70 percent of the welfare achievable under infinite GP capacity. Capacity constraints are thus not, on their own, preventing large increases in patient welfare. Whether a more sophisticated choice and priority system could achieve most of these gains given patients' incentives is an important question that we leave for future work.

V. Conclusion

This paper studies the problem of dynamically reassigning a set of agents to a set of objects when agents' preferences over objects may change over time. Our application is Norway's primary health care system, in which individuals are matched with a GP throughout their whole life, and may over time wish to switch their assigned GP. We provide direct evidence of unrealized gains from trade under Norway's current reassignment mechanism and analyze alternative mechanisms that find and execute such trades. 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-served waitlists. Patients who are matched with the least demanded GPs are especially disadvantaged. We then show that modifications to the priority system can eliminate harms to these patients while preserving more than half of the gains from TTC.

The existence of formal waitlists in our empirical setting makes unrealized gains from trade directly visible to the researcher, but similar gains could well be present in other settings where there is currently no formal way to register a desire to be reassigned. 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. Reassignment mechanisms that do exist—for example, public housing transfer systems in the United States and public school reassignment systems in New York City and Chile—often use waitlists in way that is similar to Norway's current system.

Our analysis leaves several design questions unexplored. Within the class of mechanisms we have considered, one could optimize on several dimensions we simply hold fixed—for example, the frequency with which the matching algorithm is run, the information made available to patients, or as discussed, the ability to join multiple waitlists. 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 and luck rather than by waiting times. Finally, there is a question of whether to impose formal capacity constraints (as in Norway), or to allow the market to clear through endogenous quality degradation from overcrowding (as in England). With some extension, our framework could be useful for studying these questions.

REFERENCES

Abaluck, Jason, and Abi Adams-Prassl. 2021. "What Do Consumers Consider Before They Choose? Identification from Asymmetric Demand Responses." *Quarterly Journal of Economics* 136 (3): 1611–63.

Abdulkadiroğlu, Atila, and Tayfun Sönmez. 1999. "House Allocation with Existing Tenants." *Journal of Economic Theory* 88 (2): 233–60.

Abdulkadiroğlu, Atila, and Tayfun Sönmez. 2003. "School Choice: A Mechanism Design Approach." *American Economic Review* 93 (3): 729–47.

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, 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. 2024. "An Economic Framework for Vaccine Prioritization." *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): 2291–307.

Ashlagi, Itai, Afshin Nikzad, and Philipp Strack. 2023. "Matching in Dynamic Imbalanced Markets." *Review of Economic Studies* 90 (3): 1084–124.

Baccara, Mariagiovanna, SangMok Lee, and Leeat Yariv. 2020. "Optimal Dynamic Matching." *Theoretical Economics* 15 (3): 1221–78.

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): 1749–97.

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

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

Combe, Julien, Olivier Tercieux, and Camille Terrier. 2022. "The Design of Teacher Assignment: Theory and Evidence." *Review of Economic Studies* 89 (6): 3154–222.

Combe, Julien, Umut Dur, Olivier Tercieux, Camille Terrier, and Utku M. Unver. 2022. "Market Design for Distributional Objectives in (Re)assignment: An Application to Improve the Distribution of Teachers in Schools." Unpublished.

European Observatory on Health Systems and Policies. 2023. *Norway: Health System Profile 2023*. OECD.

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): 5341–61.

Gale, David, and Lloyd Shapley. 1962. "College Admissions and the Stability of Marriage." *American Mathematical Monthly* 69 (1): 9–14.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis: Chapman & Hall/CRC Texts in Statistical Science*. CRC Press.

Gravelle, Hugh, and Luigi Siciliani. 2008a. "Is Waiting-Time Prioritisation Welfare Improving?" *Health Economics* 17 (2): 167–84.

Gravelle, Hugh, and Luigi Siciliani. 2008b. "Optimal Quality, Waits and Charges in Health Insurance." *Journal of Health Economics* 27 (3): 663–74.

Gruber, Jonathan, Thomas P. Hoe, and George Stoye. 2023. "Saving Lives by Tying Hands: The Unexpected Effects of Constraining Health Care Providers." *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, Joseph Hogan, and Fiona Scott Morton. 2017. "The Impact of Consumer Inattention on Insurer Pricing in the Medicare Part D Program." *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.

Huitfeldt, Ingrid, Victoria Marone, and Daniel Waldinger. 2026. *Data and Code for: "Designing Dynamic Reassignment Mechanisms: Evidence from GP Allocation."* American Economic Association; distributed by Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/E241067V1.

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* 132 (7): 2346–95.

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

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

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

Leshno, Jacob D., and Irene Lo. 2021. "The Cutoff Structure of Top Trading Cycles in School Choice." *Review of Economic Studies* 88 (4): 1582–623.

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

Mark, Nathaniel. 2021. "Access to Care in Equilibrium." Unpublished.

McCulloch, Robert, and Peter E. Rossi. 1994. "An Exact Likelihood Analysis of the Multinomial Probit Model." *Journal of Econometrics* 64 (1): 207–40.

Narita, Yusuke. 2018. "Match or Mismatch? Learning and Inertia in School Choice." Unpublished.

Nichols, D., E. Smolensky, and T. N. Tideman. 1971. "Discrimination by Waiting Time in Merit Goods." *American Economic Review* 61 (3): 312–23.

Norwegian Directorate of Health. 2020a. Norwegian GP Registry (Fastlegeregisteret). Helsedata. https://helsedata.no/en/forvaltere/norwegian-directorate-of-health/norwegian-gp-registry/ (accessed via approved research-data application; data available to qualified researchers via Helsedata application process).

Norwegian Directorate of Health. 2020b. Control and Payment of Health Reimbursements Database (KUHR). Helsedirektoratet. https://www.helsedirektoratet.no/statistikk-registre-og-rapporter/helsedata-og-helseregistre/kuhr (accessed via approved research-data application; data available to qualified researchers via Helsedirektoratet application process).

Norwegian Institute of Public Health. 2020. Norwegian Patient Registry (NPR). Helsedata. https://helsedata.no/en/forvaltere/norwegian-institute-of-public-health/norwegian-patient-registry-npr/ (accessed via approved research-data application; data available to qualified researchers via Helsedata application process).

Pathak, Parag, and Jay Sethuraman. 2011. "Lotteries in Student Assignment: An Equivalence Result." *Theoretical Economics* 6: 1–17.

Pereyra, Juan S. 2013. "A Dynamic School Choice Model." *Games and Economic Behavior* 80: 100–14.

Propper, Carol. 1990. "Contingent Valuation of Time Spent on NHS Waiting Lists." *Economic Journal* 100 (400): 193–99.

Propper, Carol. 1995. "The Disutility of Time Spent on the United Kingdom's National Health Service Waiting Lists." *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 Services Research* 22 (1): 1–10.

Roth, Alvin E., Tayfun Sönmez, and M. Utku Ünver. 2004. "Kidney Exchange." *Quarterly Journal of Economics* 119 (2): 457–88.

Russo, Anna. 2023. "Waiting or Paying for Healthcare: Evidence from the Veterans Health Administration." Unpublished.

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." Preprint, SSRN. http://dx.doi.org/10.2139/ssrn.3564776.

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.

Van der Vaart, Aad W. 2000. *Asymptotic Statistics,* Vol. 3. Cambridge University Press.

Statistics Norway (SSB). 2020. Population register data on Norwegian residents, 1970–2015. Statistics Norway, Microdata service for researchers. https://www.ssb.no/en/data-til-forskning/utlan-av-data-til-forskere (accessed via approved research-data application; data available to qualified researchers via SSB application process).

Verdier, Valentin, and Carson Reeling. 2022. "Welfare Effects of Dynamic Matching: An Empirical Analysis." *Review of Economic Studies* 89 (2): 1008–37.

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.