Wait Times for Surgery in the U.S.: Measurement and Allocative Efficiency in Private Insurance*

Pierre Bodéré[†], Michael J. Dickstein[‡], Guillaume R. Fréchette[§]

March 17, 2024

Abstract

In healthcare systems across the world, limited capacity implies that patients must wait to access surgical care. To evaluate the efficiency and equity consequences of rationing care via queues, however, requires comprehensive measurement of the length of these waits for multiple treatments, patient types, and insurance generosities. We employ machine learning models trained on a large claims dataset of U.S. patients with employer-sponsored insurance to measure wait times as the delay between (a) the moment our models can confidently classify a patient as in need of surgery and (b) the day of the surgery. We use this novel measure to study the distribution of wait times for roughly one million patients across many common surgeries. We find that men wait less than women, while older patients and patients with comorbidities wait longer, suggestive of potential medical inefficiencies. Similarly, we show that health insurance design affects surgical wait times in ways that may not coincide with the value of care. Using an instrument based on weekly congestion in patients insurance plan, we find that delays have adverse effects on recovery across a breadth of medical outcomes. Patients who wait a month more are 3.1% more likely to be readmitted in a hospital, spend 5.9% more, and are prescribed 6.6% more opioids in the six months following a surgery. Combining this empirical design with recent machine learning tools to recover heterogeneous effects, we quantify the medical allocative efficiency of surgical wait-lists. Applying our estimates to a subset of the surgeries that patients undergo, we find that reassigning patient priorities in the queue could substantially reduce hospital spending.

*We are grateful to Françoise Kraeber-Bodéré and Matthieu Doutreligne for numerous conversations that have substantially improved this paper. We thank Matthew Backus, Nick Buchholz, Daniel Bjorkegren, Jacob Dorn, Kate Ho, Alessandro Lizzeri, and Pietro Tebaldi for helpful comments and suggestions.

[†]Yale School of Management
[‡]NYU Stern School of Business
[§]New York University

1 Introduction

In the face of limited health care resources, waiting time often serves as a rationing mechanism in health systems across the world (Lindsay and Feigenbaum, 1984; Cullis et al., 2000; Siciliani et al., 2014). The ubiquity of wait times in healthcare delivery makes assessing their efficiency and equity consequences central to health policy discussions (Siciliani et al., 2013). However, evaluating empirically the costs and benefits of waiting for care is challenging because consistent measures of surgical wait times are difficult to observe. Few datasets contain information on surgical wait times; when wait times are recorded, their measurement typically relies on administrative markers that are difficult to compare across systems and may capture only a portion of the patient's total wait, such as hospital wait lists (Siciliani et al., 2013; Viberg et al., 2013).[1]

This paper makes four main contributions. First, we bypass the difficulty of collecting consistent wait times data from administrative sources by developing a novel approach to measuring surgical wait times from insurance claims data alone. Second, we implement this measure in one of the largest datasets of U.S. employer-sponsored insurance claims – a context where wait times information is rarely observed – and provide novel evidence on the distribution of surgical wait times for about a million patients across multiple common surgeries. Third, we propose a congestion-based instrumental variable design to evaluate the causal impact of surgical wait times on medical outcomes capturing multiple aspects of patient well-being and recovery post-surgery. Finally, by combining our measure and design with machine learning methods for the estimation of heterogeneous effects (Chernozhukov et al., 2018a; Syrgkanis et al., 2019; Athey et al., 2019), we quantify medical misallocation in surgical wait times, that is the extent to which less at-risk patients manage to access care faster.

The wait time measure we introduce consists in training machine learning models on patient medical histories and computing the delay between (a) the moment our models can classify a patient as surgical with high probability and (b) the day of the surgery itself. Intuitively, our approach can be thought of as learning medical events that are predictive of a given surgery and then using these markers to trigger wait times. A first benefit of this approach is that it does not require pre-specifying medical markers, which would necessitate time and extensive medical knowledge to scale to multiple surgeries. Second, we show that this measure performs well not only on surgeries for which a unique predictive medical marker exists but also on surgeries announced by a more complex combination of drugs, diagnoses, and procedures. Finally, this measure can be computed whenever insurance claims are available, allowing the study of wait times in contexts where they are typically not observed, such as private insurance markets, and the implementation of wait time comparisons across systems or countries.

We show that our proposed measure exhibits good medical properties. First, the trained machine

[1] As noted in Viberg et al. (2013), some OECD countries measure wait times starting from referrals—hence referral to treatment—while others instead start from patients' entry in hospital wait lists—inpatient wait time. Other countries, such as France or the United States, do not collect systematic data on wait times.

learning models achieve high classification performance and successfully identify relevant medical events predictive of each surgery. Second, our approach produces wait times that rank surgeries according to medical intuition. Urgent surgeries, such as cholecystectomy, coronary bypass, and inguinal hernia repair, have median wait times below a month, while elective surgeries, such as hip and knee replacements, have median wait times well above 6 months.[2] We show that surgical wait times typically involve multiple visits with different provider types, from primary care providers to specialists to surgeons. We also find that within surgery, there is a lot of heterogeneity in wait times: patients in the third quartile of wait times often wait five times longer than patients in the first quartile.

We investigate how patient demographics and health status explain this heterogeneity in wait times. Men wait on average 21 fewer days than women (11% standard deviation); this gap is mainly driven by elective surgeries such as cataracts and joint replacements. We also find that older patients and patients with chronic health conditions tend to wait longer, possibly reflecting the fact that these patients go through more preparatory tests before proceeding to surgery. These systematic differences foreshadow inefficiencies in wait times allocation if at-risk patients, such as those with chronic conditions, are not prioritized.

We also study the influence of health insurance design on surgical wait times. Our data provides categories for insurance plans based on the restrictiveness in access for out-of-network services, ranging from restrictive managed-care plans such as Health Maintenance Organizations (HMO) to more flexible plan types such as Premium Provider Organizations (PPO) and high-deductible plans. Rich data on claims payments also allows us to construct a measure of plans' average out-of-pocket share-the fraction of fees for services covered by patients. We find that patients enrolled in HMOs tend to wait longer than patients in more flexible plans, about a week more on average. This effect is mainly driven by elective surgeries. We also find patients enrolled in plans with higher out-of-pocket share tend to wait less. This could be due to price-sensitive patients cutting on services and thus proceeding to surgery faster or to other unobserved features of plan networks correlated with lower generosity. These results indicate that cost-sharing and network restrictions, two common cost-cutting approaches insurers use, could have opposing effects on surgical wait times.

Having computed this wait times measure, we turn to the central question of the impact of delays in access to surgical care on recovery and health outcomes post-surgery. Addressing the selection of unobservables—-stemming, for instance, from positive prioritization, whereby doctors fast-track more at-risk patients to prevent adverse outcomes—is a crucial challenge. We propose a congestion-based instrumental variable design and leverage variation within the insurance network across weeks in the waiting time associated with seeing specific provider types. This variation is due to demand and supply shocks in a patient-provider network and is, therefore, plausibly unrelated to unobserved drivers of health outcomes. We then use this variation to estimate the impact of wait times on multiple outcomes measured in the 6 months following the surgery, some previously associated with

[2] See appendix A for a detailed description of surgeries included in our sample.

delays, such as hospital readmissions, but also new outcomes, such as opioids and addictive drugs prescriptions, which shed light on the pain and decrease in quality of life associated with a worse recovery. On average, we find that wait times have moderate adverse effects on health outcomes: an increase in wait times of one month is associated with a 3% increase in hospital readmission, a 6% increase in hospital spending, and a 6.6% increase in opioid prescriptions. These effects are primarily driven by severe and urgent surgeries such as coronary bypass, inguinal hernia repair, and cholecystectomy. In contrast, for surgeries with longer median wait times, such as hip replacement, hysterectomy, and mastectomy, we find no discernible adverse effect on hospital-based outcomes. We also find that these effects persist over time, and we can still detect the adverse effects of wait times on readmission and opioids six months after the surgery took place. These adverse effects highlight the presence of medical costs when rationing access to care via wait times, and should be taken into account when comparing wait times to alternative rationing mechanisms such as prices.

Wait times for surgical care involve patients of different demographic backgrounds and varying degrees of severity of prior health conditions; as such, we anticipate these average effects to mask vast heterogeneity. We use recent advances in machine learning estimation of heterogeneous effects –specifically generalized random forests (Athey et al., 2019)– to investigate how adverse effects of wait times on hospital spending vary as a function of the high-dimensional patient health trajectory in the two years before surgery. We recover marginal effects by comparing patients with similar health trajectories but different wait times due to exogenous variation in congestion and allow these marginal effects to be functions of patient health trajectory. We propose a new calibration score suited to applications with instrumental variables to validate our heterogeneous effects models. This score compares out-of-sample the average predicted heterogeneous effects in each decile from least to most impacted patients to a double machine learning instrumental variable estimator of average effects from Syrgkanis et al. (2019) in each of these deciles. With this metric, we can credibly assess the out-of-sample validity of the marginal effects we estimate; that is, we can assert whether a new patient with a given health trajectory should be given priority to access surgery.

We find substantial heterogeneity in the adverse effects of wait times on hospital spending for 3 surgeries: coronary bypass, colectomy, and knee replacement. For coronary bypass, for instance, patients in the most impacted decile have adverse effects more than 6 times larger than the median marginal effect. These at-risk patients exhibit a worse general health status: they have higher rates of chronic conditions such as diabetes, tend to visit inpatient facilities and emergency rooms more often, take more drugs, and receive more tests. Similar profiles emerge for colectomy and knee replacement, where age is an additional aggravating factor. Finally, we show that the estimated marginal effects do not exhibit a strong negative correlation with measured wait times. This implies that more at-risk patients frequently wait longer than patients for whom wait times have little to no adverse effects. Our results, therefore, suggest that wait time allocation is inefficient for this subset of surgeries and that improving wait time targeting could generate large gains.

We implement a series of counterfactual simulations to quantify the gains from optimizing surgical wait lists. First, we estimate an auxiliary machine learning model of hospital spending, which can be combined with our marginal effects models to recover what spending would have been had a patient received the same health services but had waited longer to access surgery. Second, we reallocate wait times across patients to minimize the implied counterfactual hospital spending, holding the aggregate wait times fixed. We interpret this constraint as a way to represent a policy reorganizing existing queues without requiring healthcare providers to expand their capacity. We find that the reductions in hospital spending in the 6 months post-surgery implied by the optimal wait time allocation are large for the 3 surgeries considered: 49% for coronary bypass, 75% for knee replacement, and 26% for colectomy. Beyond these hospital spending reductions, our estimate of the adverse effect of wait times on opioid prescriptions also suggests that gains in patient quality of life could be large.

Policy Implications. Our simulation exercises uncover large discrepancies between the optimal wait times allocation and the one observed in the data in terms of implied hospital spending. However, wait times for surgical care involve multiple doctors and provider types with varying scheduling constraints. The absence of a centralized assignment system may limit how much reallocation can be achieved. In addition, patients themselves may generate additional wait time by postponing the scheduling of an appointment or canceling it altogether. As such, we do not interpret our optimal benchmark as a feasible policy target; instead, we use it to illustrate the spending and patient health gains that could result from prioritization rules informed by causal models trained on large-scale insurance claims data such as ours. In theory, prioritization improves welfare by reducing the deadweight loss from waiting times (Gravelle and Siciliani, 2008, 2009). In practice, however, designing prioritization rules that improve patient health outcomes can be challenging. Prioritization in surgical wait lists has been implemented in multiple contexts, including New Zealand, Canada, and Norway (see Siciliani et al. (2013) for a review). Common features across these applications include (1) centralized healthcare systems and (2) simple targeting rules, such as surgery-specific wait time guarantees, as in Denmark and Italy.[3] In contrast, this paper shows that there could be large gains to leveraging information from patient medical history to reorganize wait lists, not only *across* but also *within* surgeries. While the complete heterogeneous effects models rely on high-dimensional patient characteristics, we show that simpler targeting rules based only on average marginal effect by gender, age, and a commonly used index for chronic conditions severity could achieve as much as 63% of the hospital spending gains realized in the optimal allocation in the case of coronary bypass.

Related Literature. This paper contributes to several strands of literature. First, our paper belongs to a growing literature using machine learning tools to inform health policy, be it to predict

[3] Alternative prioritization strategies are implemented in Canada and Australia, where priority scores for joint replacement are assigned based on a doctor's assessment of the severity of a patient's condition.

health outcomes (Einav et al., 2018), provide benchmarks for physicians' skills (Mullainathan and Obermeyer, 2022), or investigate avenues for improving agents' decisions or physicians performance (Gruber et al., 2020; Agarwal et al., 2023). More closely related to this paper, Einav et al. (2022) use a tree-based model to construct a patient health index and measure the value-added of nursing homes. In our paper, we also summarize high-dimensional patient medical trajectory as a scalar –the probability of receiving the surgery– but focus on the evolution of this probability over time to construct our wait times measure.

Second, we contribute to the literature investigating the impact of surgical care wait times on health outcomes. Several papers from the medical literature studied the consequences of delays for coronary bypass (Sobolev et al., 2006; Sobolev and Fradet, 2008; Moscelli et al., 2016), cataract (Hodge et al., 2007) and joint replacement (Nikolova et al., 2016). These papers tend to focus on a single surgery at a time, involve relatively small samples, and do not account for patient selection, except for Moscelli et al. (2016), which also uses an instrumental variable approach based on hospital congestion. We expand on this prior research by considering multiple surgeries and new outcomes, and by estimating heterogeneous adverse effects of wait times.

Third, our research contributes to the analysis of the efficiency of the use of healthcare resources. Two recent examples related to our paper are Chan and Gruber (2020) who find that patients admitted to emergency departments are not systematically prioritized based on medical need, and Agarwal et al. (2019) who show that some matches between kidneys and patients are inefficiently lost due to the decentralized aspect of the process and inadequate incentives for hospitals to participate in the kidney exchange. We study misallocation in a different context: wait times for surgical care. This setting is also decentralized, and wait times add up to weeks or months. The type of care we consider could lead to multiple complications. These factors together likely contribute to the large inefficiencies we measure.

Overview. Section 2 describes the insurance claims data and the sample of surgeries used in the analysis. Section 3 presents our proposed measure of wait times and the methodology to compute it. Section 4 investigates heterogeneity in wait times across patient demographics and insurance plan designs. Section 5 studies the adverse consequences of wait times on health outcomes and quantifies the extent of misallocation in surgical wait times. Section 6 concludes.

2 Data Description

This section presents the insurance claims dataset used in this paper and describes the sample of surgeries and patients included in the analysis. We highlight the variables of interest in the dataset, and in particular describe the characteristics of insurance plans that are later used to evaluate the influence of insurance design on wait times. Finally, this section provides descriptive statistics on patients demographics and health care consumption by surgery.

Marketscan. This paper uses the Commercial Claims and Encounters Database from Marketscan for the years 2005 to 2013.[4] The dataset contains millions of commercially insured individuals, their spouses, and their dependents. Claims are shared by contributors which can either be large employers or health plans. In most of the analysis, we focus on claims from employers to more credibly interpret restrictions at the contributor – plan type as restrictions stemming from geographically well-defined network that can therefore impact patients' access to care. Observations are at the encounter level, and record the inpatient and outpatient medical services in addition to purchases of pharmaceutical drugs. Multiple services can take place for a given patient on a given date; we define a *visit* as patient – date pair and in most of our analysis collapse services at the visit level. To construct medical trajectories, we rely on data from three different tables linked using patient identifiers: *Inpatient Services*, *Outpatient Services*, and *Outpatient Drug Claims*. The two services tables provide detailed information on medical encounters, including diagnoses coded in the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), and procedures coded either in ICD-9-CM or Current Procedural Terminology (CPT-4) systems.[5] In addition, the data includes information on the type of doctor and medical facility corresponding to the encounter. The outpatient drug claims table provides the generic of the drug, and the purchase date that can be used to link a drug claim to a previous in- or outpatient encounter.[6] We also observe payments corresponding to the claim, broken down into deductible, copayments, and coinsurance paid by the patient, and net payments the patient's plan incurs. In addition, Marketscan distinguishes among several types of insurance plans patients can be enrolled in. Appendix table A3 reports the detailed criteria used by Marketscan to construct this classification, by we provide an overview below. The most common plan types are the following: **Premium Provider Organizations** (PPO), a form of insurance characterized by incentives for patients to use certain providers. However, PPO are still flexible in that they cover out-of-network services. **Point of Service** (POS) plans are a form of managed care that allows patients to seek out-of-network services, but may require a referral from an assigned primary care provider (PCP) –sometimes referred to as *gatekeeper*– to do

[4] The data is accessed through the National Bureau of Economic Research

[5] ICD-9-CM is based on the World Health Organization's Ninth Revision, International Classification of Diseases (ICD-9) and is the official system of assigning codes to diagnoses and procedures associated with hospital utilization in the United States. It comprises $3,824$ procedure codes and $14,025$ diagnoses codes. CPT-4 are distributed by the American Medical Association. There are more than $10,000$ codes.

[6] We link drug claims to all in- or outpatient visits that took place in the 15 days preceding the drug purchase. Unmatched drug claims are dropped.

so. **Health Maintenance Organizations** (HMO) and **Exclusive Providers Organizations** (EPO) are more restrictive forms of managed care that also assign a PCP to enrollees and that do not cover out-of-network services. Among these types of plans, PPO is considered the higher quality option, offering more flexibility but typically associated with higher premium and out-of-pocket costs. Lastly, typically younger, healthier patients may decide to enroll in high deductible plans that come in two forms: **High Deductible Health Plans** (HDHP) and **Consumer Driven Health Plan** (CDHP). In our analysis we think of plans as differentiated along two dimensions observed in Marketscan data: (1) *network restrictions*, with the most restrictive plans being HMOs and EPOs and (2) *cost-sharing* with the plans where patients bear the highest share of the costs being HDHPs and CDHPs.[7] Finally, the data also includes demographic information on patients, including sex and age. Marketscan has two appealing features for the purpose of this study: first, the data covers millions of lives in a longitudinal fashion, providing large sample of patients for many surgeries with their corresponding medical trajectories leading up to the surgery. Second, Marketscan mostly consists in employed individuals, thus offering a window in the health care uses of non-elderly patients in a broad range of ages.

Surgery Definitions and Sample Selection. We select 17 surgeries for which to define wait times based on two criteria: included surgeries should be frequent enough to allow for training statistical models and should be therapeutic rather than diagnostic so that the surgery can be considered the endpoint of a care episode.[8] The sample includes the universe of patients in Marketscan who have received one of these surgeries between 2007 and 2013 and are continuously enrolled at least two years prior and six months after their surgery. In addition, we also sample a *control group* from the set of patients having received *none* of the selected surgeries for the purpose of training the machine learning models used to construct our wait time measure.[9] We provide brief descriptions of the surgeries included in the analysis from the U.S government website MedlinePlus, a service of the National Library of Medicine.[10]. Appendix tables A1 and A2 provide a description of the surgeries included in the analysis as well as a list of the procedure codes used to identify surgery recipients in Marketscan. Table 1 shows the number of patients in each surgery and in the control group, and their corresponding demographics. A striking benefit of Marketscan is the size of the sample: for most of the surgeries, we can identify dozens of thousands of patients that satisfy our pre and post-surgery enrollment criteria. The third column of Table 1 reports patients' Charlson comorbidity index, a measure of general health condition known to be a good predictor of

[7] In practive plans also differ in terms of the premiums they charge, and network details such as the specific hospitals or the quality of doctors, but this information is not available in Marketscan.

[8] In future iterations of this paper, we plan on including more surgeries that could fit our criteria, such as bariatric surgery. Note that in Section 3 we single out three surgeries among the 17 for which our method to compute wait times is not well-suited; this is because either samples are too small or the surgery is an emergency, so that their is no wait time.

[9] We sample exactly $100,000$ potential control patients for each of the surgery years $2007 - 2013$.

[10] The website and its search function used to recover the surgeries descriptions can be accessed at https://medlineplus.gov

mortality in longitudinal studies Charlson et al. (1987); Quan et al. (2005).[11] Charlson indices rank control and surgical patients according to intuition: control patients have an average index of 0.19 suggesting that serious commorbidities are rare in that group. Patients in elective surgeries such as cataract, hip and knee replacement also tend to have low Charlson indices.c In contrast, patients waiting for heavier surgeries such as mastectomy, kidney and liver transplants have high average Charlson indices (3.27, 2.75 and 8.01 respectively). The surgeries included in our sample vary in the demographics of patients that receive them: some such as hysterectomy, mastectomy or prostatectomy are overwhelmingly received by either women or men. Some surgeries such as cataract, knee replacement and endareterectomy are mostly performed on older patients aged between 55 and 64, while for others such as hysterectomy, cholecystectomy or colectomy the share of recipients below 55 is higher. Table 2 reports statistics of patients health care utilization. Recipients of heavier surgeries consume more health care across the board: for instance, the mean number of inpatient visits in the year of the surgery for liver transplants is 32, and the average inpatient total payments in that year amount to $302,100, while cholecystectomy recipients only go through 3 inpatient visits on average and generate $9,200 of inpatient spending. While for most surgeries, inpatient payments dwarf outpatient ones, for others the opposite is true, reflecting patient trajectories where most of the care happens in outpatient setting: this is the case or cataract, ($3,200 inpatient and $12,800 outpatient), inguinal hernia ($2,200 v.s. $10,000) and mastectomy ($9,100 v.s. $51,600). These differences highlight the importance of taking into account both inpatient and outpatient care when constructing patients' trajectories for the purpose of this paper. Indeed, the relevance of outpatient versus inpatient information in determining how likely a patient is to receive a surgery might vary across surgeries, but also across patients depending on the insurance they are enrolled in and the type of care settings they have access to.

[11] More precisely, we use the package developed by (Gasparini, 2018) to compute morbidity weights for 17 groups of medical conditions: Myocardial infarction, Congestive heart failure, Peripheral vascular disease, Cerebrovascular disease, Dementia, Chronic pulmonary disease, Rheumatologic disease, Peptic ulcer disease, Mild liver disease, Diabetes without chronic complication, Diabetes with chronic complication, Hemiplegia or paraplegia, Renal disease, Any malignancy, including leukemia and lymphoma, Moderate or severe liver disease, Metastatic solid tumor, AIDS/HIV. These conditions receive a score from 1 to 6 based on associated morbidity, and a patient's Index is just the sum over the encountered conditions.

Table 1. Sample Demographics

				Share in Age Bin				
Surgery	Number Patients	Charlson Index	Share Women	[0-17]	[18-34]	[35-44]	[45-54]	[55-64]
Appendectomy	139,829	0.47	0.51	0.24	0.28	0.17	0.18	0.13
Cataract	228,013	0.82	0.55	0	0.01	0.03	0.2	0.76
Cholecystectomy	348,504	0.78	0.73	0.02	0.21	0.23	0.28	0.26
Colectomy	61,725	2.23	0.51	0.01	0.06	0.13	0.33	0.46
Coronary Bypass	40,046	1.16	0.2	0	0	0.04	0.26	0.69
Endarterectomy	9,465	1.11	0.39	0	0	0.01	0.17	0.82
Hip Replacement	61,707	0.68	0.49	0	0.02	0.07	0.31	0.6
Hysterectomy	198,045	0.72	1	0	0.09	0.39	0.39	0.13
Inguinal Hernia	112,881	0.46	0.1	0.14	0.1	0.14	0.27	0.35
Kidney Transplant	4,421	2.75	0.39	0.04	0.12	0.19	0.31	0.34
Knee Replacement	127,998	0.68	0.6	0	0	0.03	0.26	0.7
Liver Transplant	1,391	8.01	0.33	0.05	0.06	0.08	0.29	0.53
Mastectomy	84,673	3.27	0.97	0.01	0.04	0.16	0.37	0.42
Neprhostomy	248	1.55	0.51	0.08	0.13	0.14	0.28	0.37
Prostatectomy	32,482	2.61	0	0	0	0.02	0.26	0.72
Splenectomy	4,342	2.43	0.53	0.1	0.17	0.14	0.26	0.34
Thyroidectomy	31,451	2.05	0.82	0.01	0.13	0.22	0.33	0.31
Control	700,000	0.19	0.51	0.26	0.2	0.17	0.21	0.17

Notes: The table shows demographics for surgical and control patients. For surgeries, the sample consist in all surgery recipients between 2007 and 2013 that satisfy the following conditions: they have not received the same surgery before, are enrolled 2 years before and 6 months after the surgery, and have their pharmaceutical drugs reported. Charlson Index is computed at the last visit before surgery for surgical patients, and is just averaged over the whole trajectory for control patients.

Table 2. Sample Health Care Utilization

	Inpatient Visits				Inpatient Pay ($1000)				Outpatient Visits				Outpatient Pay ($1000)			
Surgery	Mean	25th	50th	75th	Mean	25th	50th	75th	Mean	25th	50th	75th	Mean	25th	50th	75th
Appendectomy	2	1	1	2	12	0	7	14.8	12	4	8	15	10.1	2	6.6	13.7
Cataract	2	1	1	1	3.2	0	0	0	19	9	14	24	12.8	5.6	8.4	13.5
Cholecystectomy	3	1	1	2	9.2	0	0	10	18	8	13	22	14.4	7.2	11.4	17
Colectomy	9	2	5	11	52.1	19.8	30.8	53.4	28	12	20	37	24	4.7	9.6	23.1
Coronary Bypass	11	6	8	12	81.5	45	63.4	93.4	30	15	25	40	16.7	5.3	10.3	18.5
Endarterectomy	5	1	2	6	30.9	11	17.3	32.9	24	12	19	30	14.3	4.5	8.6	16.6
Hip Replacement	4	1	3	4	35.3	21.8	28.5	40.2	31	17	27	40	10	3.5	6.2	11
Hysterectomy	2	1	1	2	10.8	1.9	7.4	13.8	17	9	13	21	12.5	3.4	8.1	15.7
Inguinal Hernia	1	1	1	1	2.2	0	0	0	12	5	8	15	10	5	7.4	11.1
Kidney Transplant	11	5	7	12	108.7	43.3	88.9	137.9	76	42	62	94	72.4	22.9	46	89.2
Knee Replacement	4	2	3	4	32.9	20.4	27.3	39	38	24	34	48	11.5	5	8.1	13.6
Liver Transplant	32	10	20	42	302.1	160.6	238.5	368.4	59	39	55	73	54.9	21.3	39.1	66
Mastectomy	2	1	1	1	9.1	0	0	11.7	40	20	35	55	51.6	19.2	39.6	68.7
Neprhostomy	6	1	3	6	35.2	7	19.7	42.7	31	15	23	37	31.3	8.9	18	34.5
Prostatectomy	2	1	1	2	20.1	11	16.4	23.3	22	13	18	26	13.4	4.6	7.9	14.6
Splenectomy	14	2	7	16	88.8	20.5	41.1	94.8	32	13	25	43	30.4	5.8	14.9	36
Thyroidectomy	2	1	1	1	7.6	0	0.7	10.2	25	15	21	30	18.3	8.6	14.5	22.6
Control	1	1	1	1	0.7	0	0	0	7	1	4	9	2.3	0.2	0.6	2

Notes: The table shows statistics of inpatient and outpatient healthcare utilization for surgical and control patients. A visit is defined as patient - day pair, so that if several services are performed on the same day, they only count as one visit. For surgical patients, visits and payments are computed in the calendar year of the surgery. For control patients, statistics are computed in the year in which they're sampled among the population of non surgical patients.

3 Detection to Treatment: a Data-driven Measure of Wait Time to Surgery

This section proposes a novel measure of wait times for elective surgery. The availability of large insurance claims datasets opens up the possibility to rely on medical data to learn relevant medical events marking the starting date of wait time. We implement this idea by training surgery-specific classifiers to distinguish future surgery recipients from control patients. The fitted models achieve high classification performance and gives sensible results. Indeed, it detects surgical patients earlier for less emergent procedures, and the medical events that are identified as important triggers of increased probability are sensible (given the surgery).

Defining Wait Times. While the end date of the wait time for surgery is unequivocal and coincides with the moment the procedure is received, what the relevant starting point for waiting time is less obvious (DeCoster et al., 1999; Viberg et al., 2013). Among the countries that do collect systematic information on wait times for surgery, starting dates vary depending on the specifics of the country's health care system (Viberg et al., 2013). Commonly used starting dates include the moment a referral to a specialist is written by a primary care physician, the moment the decision to treat via surgery is made, and the moment a patient enters an inpatient waiting list at a facility. Each of those waiting times may be of interest, but they preclude comparisons, and only highlight one small part of the overall experience of the patient. Our measure will be more encompassing and allow for comparisons, in the case of the current paper for instance, analysis from patients using different insurance plans and who resides in different states. In addition, such data is essentially unavailable outside of single-payer public systems. In the absence of administrative data on wait times, public health and medical researchers have resorted to alternative measures relying on choosing as a starting date a specific test or procedure announcing the upcoming surgery. Examples include optical biometry for cataract surgery, as in a recent study by Chen et al. (2021). This approach is well-suited when working with only one surgery and when there exists such a well-defined marker present in the majority of surgical patients' trajectories.[12] However, even when such highly predictive events exist; patient's medical trajectories are so varied that such identification misses many cases. Furthermore, this selection is not random, in many cases the trajectory is itself related to wait time. Our measure of wait times consists in a statistical generalization of the "optical biometry" as predictive of cataract approach, that allows for rich and diverse trajectories. If, given a surgery recipient's medical history, the procedures, diagnostics or drugs received at a given visit imply with high probability that the patient will receive a surgery in the future, we set that visit as our starting date for the wait time. This measure provides two main advantages over earlier approaches. First, it relies on medical data only and thus abstracts from the arbitrary

[12] A notable exception is DeCoster et al. (1999) who study wait times for 10 elective surgeries using as a trigger for wait times the preoperative visit to the operating surgeon. While closer in spirit to our measure, this approach also leaves out patients without preoperative visits and might miss the portion of wait time between the moment the decision to proceed to the surgery was made and the preoperative visit.

administrative checkpoints a given system chooses to record. It can be applied to multiple countries or insurance plans, and therefore enables systemic comparisons. Second, our measure can be applied to multiple surgeries at once without requiring extensive medical expertise. The statistical nature of our definition relaxes the requirement that a single marker appears on all surgery recipients' trajectories: it can accommodate different pathways to surgery depending on a patients' medical histories. We highlight these properties of our approach later in this section, after having described our measure and the method to compute it. The next paragraph introduces notation to formalize this definition.

Notation. Insurance claims data systematically collect information on medical services received by patients for reimbursement purposes. Consider a patient i covered by insurance. Typically, a claim for a visit occurring at time τ records the procedures performed on patient i, $\{p^1_{i\tau}, \ldots, p^{n^p}_{i\tau}, \} = P_{i\tau}$ with each procedure p^n coded as either a CPT-4 code or and ICD code. Visit claims also contain diagnoses received $\{d^1_{i\tau}, \ldots, d^{n^d}_{i\tau}, \} = D_{i\tau}$, and information on the type of provider encountered, whether the setting was in or outpatient, and potentially additional information which we denote as $Z_{i\tau}$. In addition, we also observe prescription drugs purchased in outpatient pharmacies by patient i following visit τ which we denote by $M_{i\tau}$. Procedures, diagnoses, drugs and the characteristics of visits associated with patient i up to time t constitute the patient's observed medical history $\mathcal{H}_{it} = \{P_{i\tau}, D_{i\tau}, M_{i\tau}, Z_{i\tau}\}_{\tau \leq t}$. Consider a surgery s and define the probability that patient i receives the surgery s at any point in the future given i's medical history at t as $\mathbb{P}(s|\mathcal{H}_{it})$. Given a confidence threshold $\bar{p}$, and denoting by $\mathcal{T}_i$ the set of dates of visits for patient i, we propose as the starting date of wait time the first visit date the probability of surgery exceeds the confidence threshold:

$$t^d_{is}(\bar{p}) = \min\{t \in \mathcal{T}_i, \mathbb{P}(s|\mathcal{H}_{it}) \geq \bar{p}\} \tag{1}$$

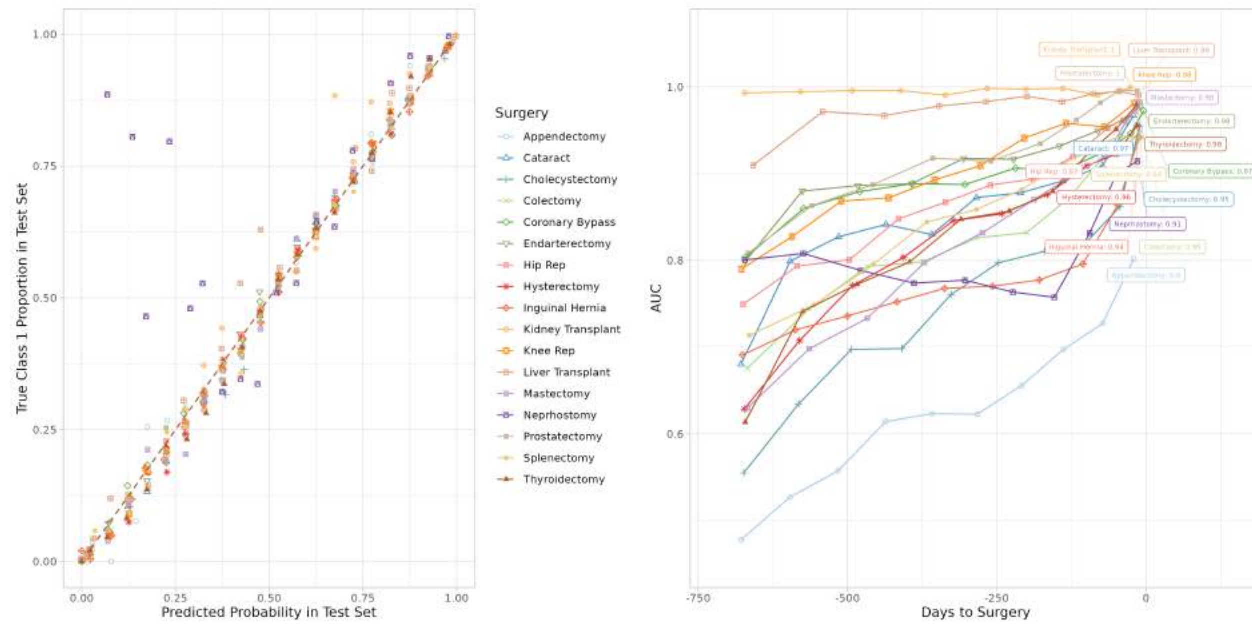
Wait time to surgery then consists in the difference in days between this detection date $t^d_{is}(\bar{p})$ and the day of the surgery t^s_{is}:

$$w_{is}(\bar{p}) = t^s_{is} - t^d_{is}(\bar{p}) \tag{2}$$

Implementation. In practice, $\mathbb{P}(s|\mathcal{H}_{it})$ is unknown. We estimate this probability using a classifier trained to distinguish recipients of surgery s from a control group of randomly sampled non-surgical patients. The estimated classifier provides a mapping from medical histories to the probability of receiving surgery s in the future, $\mathcal{H}_{it} \rightarrow \hat{p}_{ist} \in [0, 1]$. We use a gradient boosting classifier and represent a patient's medical history as a matrix with as many rows as patients' visits, and as many columns as procedure, diagnoses and drug codes.[13] Each element of this history ma-

[13] As show in Appendix B, our gradient boosting model achieves higher classification performance than a LASSO-Logistic Regression trained on the same variables, suggesting that *interactions* of medical events provide information predictive of

trix is an indicator for whether a code has been observed in that visit or before. We have explored an alternative approach to encoding histories include recurrent neural networks, but as illustrated in Figure 1 we are able to achieve excellent classification performance with our comparatively less computationally intensive model.



(a) Calibration

(b) AUC Over Time

Figure 1. Model Classification Performance over Time

Notes: Panel (a) shows the fraction of patients receiving the surgery in the test set by bin of model predicted probability. Panel (b) shows the average AUC scores at various points in patients' medical trajectories before surgery. The scores are computed on a hold-out sample of treated and control patients. Treated patients visits are split in deciles of how far back in time they occurred before the surgery, and control patient visits are just randomly selected. AUC are computed over each of these deciles.

Our definition of wait times requires to interpret the scores $\hat{p}_{ist}$ predicted by the trained models as probabilities of future surgery. The left panel of Figure 1 shows that our trained classifiers are well calibrated, providing support for this interpretation of classifiers' scores as probabilities.[14]

The right panel of Figure 1 shows how the Area Under the Curve (AUC) on a hold-out test sample of our trained model evolves over time.[15] First, when the model has access to the entire medical history up to the last visit before the surgery, it achieves excellent classification performance. For almost all the surgeries considered, except Appendectomy and Nephrostomy, the final AUC is above 0.9, and for the majority it is above 0.95. This reflects the fact that in the days close to the surgery the vast majority of surgical patients have received diagnoses or procedures that are highly predictive of the surgery ahead and rarely observed for non-surgical patients. Second, the right

future surgeries.

[14] In the context of classification tasks in machine learning, calibration refers to the extent to which the predicted class probabilities reflect the true class probabilities.

[15] The AUC ranges from 0.5 (random) to 1 (perfect classifier) and is the area under the locus of False Positive (FP) and True Positive (TP) rates as the threshold of probability above which an observation is considered positive varies.

panel of Figure 1 shows that the model can better predict surgeries as there is a longer medical history. This is consistent with medical events getting more predictive over time and facilitating better classification. The levels of these performance graphs foretell the wait times associated with each surgery: for instance kidney and liver transplants, which both involve long wait times and specific procedures such as dialysis, exhibit from the very beginning near perfect classification performance. This implies that the detection date $t_{is}^{d}(\bar{p})$ for these surgeries is likely to occur early in patient's medical history, resulting in longer wait times according to definition 2. Finally, as a falsification we include in our sample Appendectomy, an emergent surgery that should be very difficult to predict in advance as few early signs exist. As shown in the middle panel of Figure 1, classification performance for Appendectomy remains very poor (AUC around $0.6 - 0.7$) until the very last days before the surgery takes place, reflecting this absence of early signs. For that reason, appendectomies are not included in our subsequent analysis. Figure C11 in Appendix displays the true positive and false positive rates graphs resulting from the classification over patients' entire medical histories. While definition 1 in principle allows both for surgical patients without any associated detection time and for control patients classified as surgery recipients, Figure C11 shows that in practice this is not a concern. Even with a detection threshold as high as $\bar{p} = 0.9$, the classifiers correctly detect about 90% of surgical patients and incorrectly predict as surgical less than 5% of control patients for almost all surgeries considered. Finally, while the choice of $\bar{p} = 0.9$ might seem arbitrary, Figure C12 in Appendix shows that wait times implied by different thresholds are highly correlated for the vast majority of surgeries in the sample.

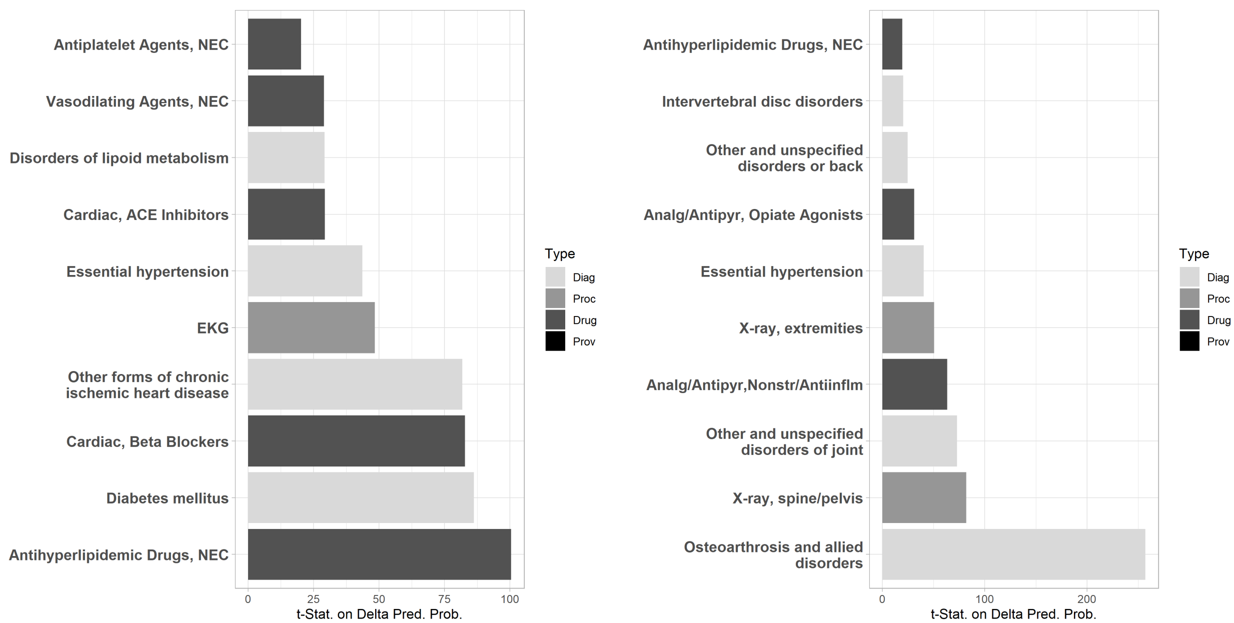
Interpretation of the Estimated Models. The definition of wait time outlined above can be interpreted as a data-driven way to detect medical events predictive of a future surgery and use these events as a starting date for wait time. To illustrate this point, we investigate which elements of patients' medical histories generate large increases in predicted probabilities: these medical events are triggers likely to lead to a patient being detected as surgical, and play the same role as the optical biometry procedure use to compute wait times in Chen et al. (2021). The following regressions shed light on the triggers learned by the trained classifiers:

$$\Delta\hat{p}_{ist} = \sum_{e} \alpha^{(e)} \Delta\mathcal{H}_{it}[t, e] + \epsilon_{ist} \tag{3}$$

where $\Delta\hat{p}_{ist}$ is the change in predicted probability of surgery s for given patient i between a visit at t and patient i's previous visit, and $\Delta\mathcal{H}_{it}[t, e]$ stands for the change in the element of patient i's medical history encoding whether the medical event e was received at or prior to time t. The coefficients of interest $\alpha^{(e)}$ capture how much receiving a procedure, drug, or diagnosis e moves the predicted probability of receiving the target surgery s. We show in Appendix B that these regressions highlight similar variables than the more computationally intensive permutation importance typically used to interpret machine learning models, with the additional benefit of providing the direction in which a variable influences predicted probabilities. Figure 2 displays the most important predictors of change in probabilities, consisting in the medical events with the highest absolute t-statistics $\frac{\hat{\alpha}^{(e)}}{\hat{s.e.}(\hat{\alpha}^{(e)})}$ in the regression above, for a subset of surgeries. This subset is selected for brevity, but the reader may find the same graphs for the remaining procedures in Appendix Figures C13, C14. Figure 2 shows that for each of the selected surgeries, the trained classifiers are able to identify medically relevant drugs, procedures, and diagnoses. For instance, antihyperlipidemic and beta blocker drugs increase the probability of receiving a coronary bypass, and osteoarthrosis diagnosis is predictive of hip replacement. The differences between the profiles of variable importance illustrate that some surgeries have clear markers, such as osteoarthrosis diagnoses for hip replacement, while other surgeries like coronary bypass are predicted by more complex patient histories combining a mixture of drugs, latent diabetes or hypertension, and possible heart problems. Relative to using a single procedure or diagnosis as a starting point for wait times, a benefit of our measure is that it can accommodate these two types of surgeries. To illustrate this property, we compare our method to a deterministic measure of wait times corresponding to the difference between (a) the first visit at which the most predictive medical event is observed and (b) the day of the surgery.[16] For these two approaches, we first compare the fraction of surgical patients with an assigned wait time (true positives, TP) and the fraction of control patient incorrectly assigned a wait time (false positives, FP). When a highly predictive medical event exists, such as for inguinal hernia or joint replacements, as seen in Figures 2 and C13:C15, the two approaches perform equally well, as reflected in the small differences in true positive and false positive rates in Table C4. In fact, as shown in Figure C16, the wait times recovered from the two approaches are highly correlated for surgeries with predictive events, consistent with the interpretation of our method as learning surgical markers and starting the wait time when they first occur. In contrast, for surgeries with more complex patterns, such as coronary bypass, mastectomy or hysterectomy, our approach performs better than the deterministic one: as shown in Table C4, detection to treatment tends to achieve high true positive rates, and is less likely to return false positives than the measure based on the most predictive event. For these more complex procedures, detection to treatment tends to return shorter wait times as shown in Figure C16, consistent with the model using a combination of predictors rather than relying on just one.

Discussion. Patients' interactions with healthcare systems leading them to surgery involve multiple steps. From the initial visit to a primary care physician following a developing health issue to the final pre-operative visit with a surgeon, one can measure aspects of wait time from different starting points. The earlier the starting date, the more downstream sources of delays a measure encompasses. For instance, inpatient waiting time starting at the inclusion of patients in facilities' wait-lists only captures facility capacity constraints, while referral to treatment also incorporates potential delays due to difficulties in scheduling specialist visits. The measure proposed in this section starts the wait time as soon as a patient's medical history shows with high probability that

[16] We define the most predictive event as the medical event the highest t-stat in regression 3



(a) Coronary Bypass

(b) Hip Replacement

Figure 2. Delta Probability Variable Importance

Notes: We illustrate the key medical "events" that underlie our prediction of surgical need for two procedures: coronary bypass and hip replacement. In each panel, we list the top 10 events—including procedures, drugs taken, and diagnoses— that suggest the patient will receive the focal surgery in the future. The events are selected and ordered in the figure using the t-statistic we find from a regression of (a) the change in predicted probability of surgery between two visits, $\Delta\hat{p}_{ist}$, on (b) the presence of the event between these visits.

doctors have identified a medical condition requiring surgery. As such, it includes any source of delays from that detection point onward: hospital capacity constraints, but also wait time implied by additional visits which could stem from further diagnosing or required preoperative testing.

4 Wait Times Heterogeneity Across Patients and Insurance Plans

We use the measure defined in the previous section and show that it produces medically reasonable wait times, shorter for emergent surgeries and longer for joint replacements and organ transplants. We investigate heterogeneity in wait times by patient and insurance plan characteristics. Women, older and more co-morbid patients wait longer across all surgeries. Insurance plan design impacts wait times, in particular for elective surgeries: patients in managed care plans with out-of-network restrictions wait longer, and, conversely, patients in plans with a higher degree of cost-sharing wait less. We also find that insurance designs assigning a PCP to patients acting as a gatekeeper lead to longer wait times spent between PCP visits. Our results highlight a trade-off between reducing costs through network restrictions and timely access to surgery.

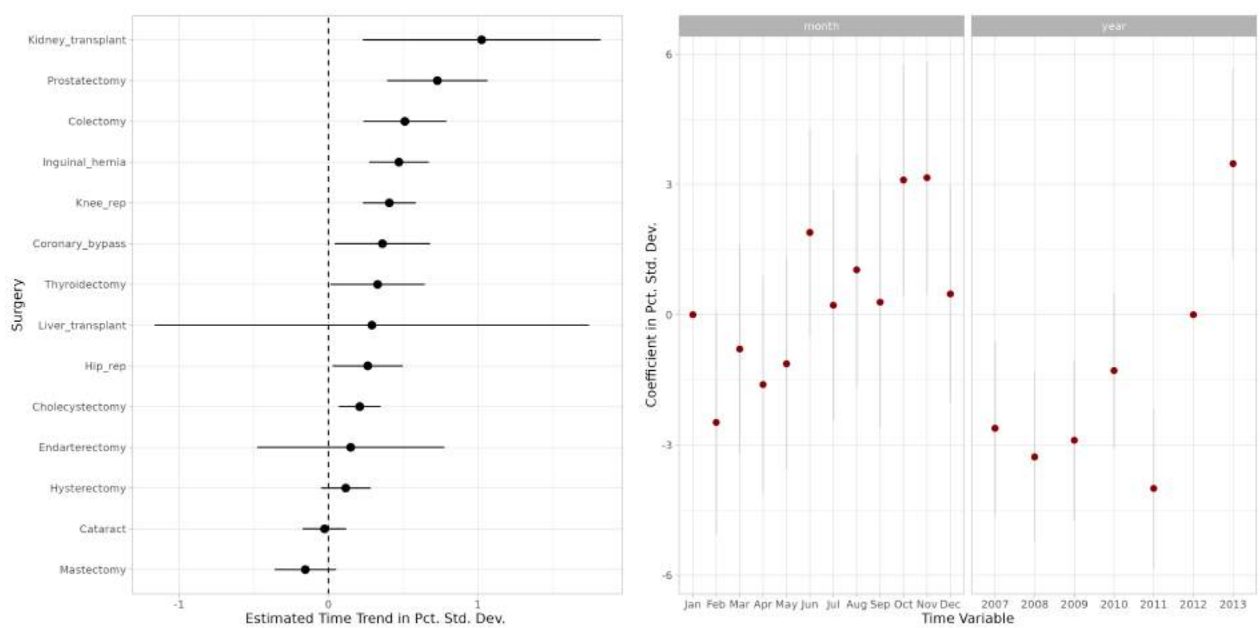
Wait Times Distributions. Table 3 shows the 10th, 25th, 50th, 75th and 90th percentiles of wait times at $\bar{p} = 0.90$ detection threshold for each surgery in our sample. The measure of wait times introduced in the previous section ranks these surgeries according to what medical intuition would suggest. More emergent procedures such as cholecystectomy, coronary bypass and inguinal hernia repair have median wait times below a month. At one extreme, appendectomy, which is typically performed immediately after an acute appendicitis diagnosis is made due to concerns about increased postoperative morbidity resulting from delays (Shin et al., 2014) exhibits a median wait time of just one day. At the other extreme, median transplant surgery recipients wait more than 600 days for a liver and about 700 days for a kidney.[17] In between, median wait times range from one to two months for mastectomy and cataract surgery to slightly less than a year for knee replacement. Table 3 also reveals substantial heterogeneity in wait times within surgery: for most procedures, the inter-quartile range is large and about 10% of patients wait more than 500 days. Consistent with Agarwal et al. (2021), we find increasing trends in wait times for kidney transplants over our sample period of about 1 percent standard deviation a year, as shown in the left panel of Figure 3. Most surgeries in our sample display increasing trends, although the magnitudes are small and the overall trend across all surgeries is quite noisy, as shown in the right panel of Figure 3. In the remaining of this section, we investigate how wait times vary across patients characteristics and insurance plan designs.

[17] Our wait time measure is capped at 720 days for data availability reasons, which might be binding for transplant surgeries, resulting in a downward bias. Agarwal et al. (2021) document average times on wait list to receive a kidney of over 2 years.

Table 3. Distributions of Wait Times by Surgery (days)

Surgery	Pct 10th	Pct 25th	Pct 50th	Pct 75th	Pct 90th
Appendectomy	1	1	1	1	43
Cholecystectomy	2	6	17	49	191
Coronary Bypass	2	5	24	396	621
Inguinal Hernia	6	13	27	57	160
Mastectomy	14	23	40	101	324
Cataract	12	23	55	300	534
Colectomy	6	18	69	258	515
Endarterectomy	6	23	91	468	636
Neprhostomy	11	32	104	433	650
Hysterectomy	18	42	108	341	540
Prostatectomy	57	81	124	287	552
Thyroidectomy	34	61	127	381	610
Splenectomy	7	40	173	458	649
Hip Replacement	34	80	224	466	625
Knee Replacement	41	120	352	575	676
Liver Transplant	187	414	612	691	712
Kidney Transplant	427	625	699	716	720

Notes: Wait times with detection threshold at $\bar{p} = 0.9$



(a) Trends by Surgery

(b) Month and Year

Figure 3. Trends and Seasonality in Wait Times

Notes: Panel 3a shows linear trends by surgery on standardized wait times estimated with months fixed effects. Panel 3b shows month and year fixed effects on standardized wait times in a regression re-weighted so that all surgeries are contribute equally. Confidence intervals are based on robust standard errors.

Investigating Heterogeneity in Wait Times. We evaluate systematic differences in wait times across patients and insurance plans using a series of regressions of the following form:

$$w_{ispt} = \boldsymbol{x}_i'\boldsymbol{\beta_x} + \boldsymbol{y}_{pt}'\boldsymbol{\beta}_p + \phi_{st} + \epsilon_{ispt} \tag{4}$$

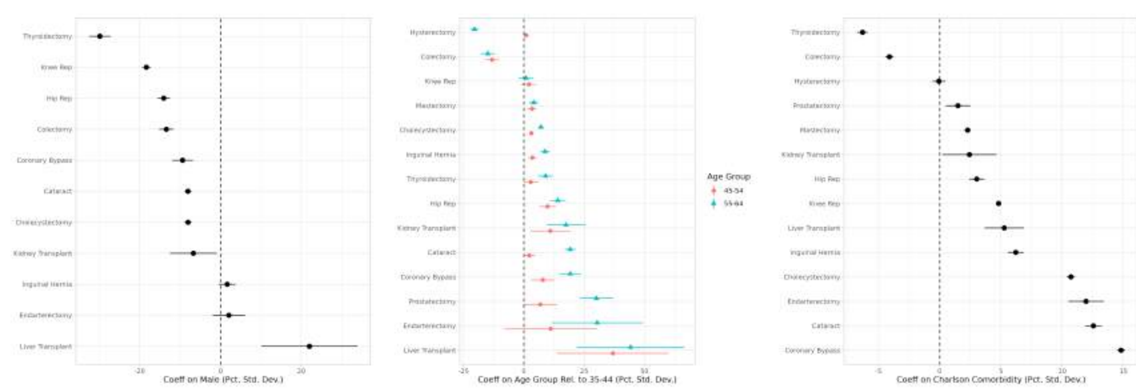
where $\boldsymbol{x}_i$ is a vector of fixed patient demographics and $\boldsymbol{y}_{pt}$ is a vector of insurance plan characteristics that we describe in details below, and ϕ_{st} denotes surgery–year and surgery–month fixed effects capturing surgery specific trends and seasonalities described in the previous paragraph. In theory, patients may select into one of the various plans proposed by their employers based on information –unobserved to the econometrician– regarding their expected future healthcare utilization which may bias the coefficients of interest $\boldsymbol{\beta_p}$. To address this concern, in addition to ordinary least squares we also compute instrumental variable estimators of $\boldsymbol{\beta}_p$ plan characteristics are instrumented by their average in a patient's employer e. While plan choice within employer may be correlated with unobserved health, we assume that employer choice is not.[18] In practice, we find that our instrumental variables estimates tend to align with the ordinary least square ones in direction and magnitude. We describe our in more details in the paragraph where we present the estimated coefficients for insurance plan characteristics. In addition to pooled regressions across surgeries, we also investigate heterogeneity in wait times by running these regressions surgery by surgery. We first describe the estimated coefficients on patient demographics before moving on to differences in wait times induced by insurance design.

Wait Times and Patient Demographics. Panel A in Table 4 shows coefficients for age, gender, and Charlson comorbidity score status from model 4.[19] Men wait about 11% standard deviation (22 days) less then women, and the oldest group in our sample of working age patients (55-64) has wait times 4.8% standard deviation (8 days) above patients aged between 35 and 44. This inequality in wait times penalizing women and older patients has been documented for cataract and hip replacement surgery in some contexts (Smirthwaite et al., 2014; Hacker and Stanistreet, 2004), although findings vary across countries (Landi et al., 2018). Moreover, we show that orthopaedics and ophthalmology are not isolated cases: as displayed in the left and middle panels of Figure 4, older patients and women wait more for almost all the surgeries in the sample. Disparities between men and women have been documented for a range of health outcomes (Cabral and Dillender, 2021). Several mechanisms could explain longer wait times for women: delays could results from providers misdiagnosing women and thus being reluctant to prescribe more aggressive care. Women could also be less likely to consult male specialists for certain types of care (McDevitt and Roberts, 2014). Alternatively, shorter wait times for men could be the consequence of a lower

[18] While not entirely satisfactory, we think of this assumption and of the resulting instrument as a way to remove the selection from one of the margins of choice. Employer choice involves many different considerations in addition to health benefits.

[19] The Charlson comorbidity score measures patient health status by assigning weights to diagnoses such as diabetes, AIDS, or cancer based on the morbidity associated with each condition. It has been shown to be a good predictor of mortality in longitudinal studies.

propensity to consume preventive care (Vaidya et al., 2012) and a tendency to postpone interactions with providers until their medical condition is advanced, resulting in shorter wait times. Finally, differences in willingness to wait due to men having less flexible work schedules than women could also explain this discrepancy (Dunn, 2018). Panel A in Table 4 also shows that patients with a higher Charlson comorbidity score—reflecting a worse general health condition—wait longer than healthier patients. This correlation is found across almost all surgeries except thyroidectomy, colectomy and hysterectomy, including for serious, emergency procedures such as coronary bypass and cholecystectomy is shown visually in Figure 4c. On the one hand, this positive correlation is puzzling: from the perspective of a wait list designer optimizing post-operative recovery, it would be optimal to allocate more at-risk patients faster. Nevertheless, more co-morbid patients have been found to wait more in the context of bariatric surgery (Alvarez et al., 2019) and for in-hospital surgical queues for hip replacement (Hamilton et al., 2000; Wei et al., 2019). On the other hand, this can result from co-morbid patients requiring more preparation before surgery, or from providers being wary of subjecting patients in poorer health condition to aggressive treatments. We evaluate the medical efficiency of the wait times allocation in the next section.



(a) Gender (b) Age (c) Charlson Comorbidity Score

Figure 4. Wait Times across Demographics

Notes: Each panel shows the coefficients and 95% confidence intervals on a given demographic variable estimated in isolation from other patient demographics, surgery by surgery, including month and year of surgery fixed effects. Standard errors are robust.

Wait Times and Health Insurance Design. Marketscan provides information on the type of plans in which patients are enrolled. There are two main dimensions of differentiation across plans which could in principle influence wait times: *network restrictions*, which limit access to certain services or providers in a patient's area, and *cost-sharing*, which requires patients to cover out-of-pocket (OOP) a higher share of health care spending. The *plan type* variable in Marketscan is constructed based on a combination of network restrictions and cost-sharing criteria. In the following analysis, we therefore present first the effects of various plan types on wait times, and second the effects of cost-sharing and plan types controlling for cost-sharing separately. An overview of the main plan types and their characteristics can be found in Section 2. We provide a detailed description of the criteria used by Marketscan to construct the plan types in Appendix table A3. We study how wait times vary across plan types by first including these plan types as the only plan characteristics regressors $\boldsymbol{y}_{ipt}$ in regression 4, while still including gender, age and Charlson index as demographic controls and surgery–year and surgery–month fixed effects. Columns (OLS1), (OLS3), (IV5), and (IV7) of Panel B in Table 4 show the coefficients on plan types estimated from ordinary least squares (OLS) and instrumental variables (IV) regressions, expressed in levels and in percentage of standard deviations,for a sample focusing on employer contributors only.[20] The plan type excluded from the regression is PPO. First, across estimators, we find that more restrictive managed care plans such as HMO and capitated POS lead to longer wait times than the more flexible PPO plans. The estimated delay with IV excluding health plan patients (column 5-7) suggests that patients in HMOs wait 6.6% of a standard deviation (12.6 days) more than similar patients expecting to receive the same surgeries in PPOs. Figure C17a in Appendix, shows the impact of HMOs relative to PPOs by surgery. The average effect in Panel BTable 4 is driven primarily by elective surgeries, such as cataract and joint replacement surgeries, where the estimated coefficients range between 5% to over 15% of a standard deviation. The effect is close to null for emergencies like coronary bypass, cholecystectomy or organ transplants. In contrast, we find that patients in POS plans wait about 4.1% of a standard deviation (7.4 days) less than similar patients in PPOs. Similarly, patients in HDHP plans wait less, with an estimated difference with PPOs of about 3.9% of a standard deviation (8 days). The estimated effects of HMO and HDHP designs in particular suggest that two approaches meant to rein in healthcare spending have opposite consequences on wait times. On the one hand, plan types involving network restrictions such as HMO and capitated POS are associated with longer wait times. On the other hand, plan types characterized by a high degree of cost sharing such as HDHP and CDHP imply shorter wait times. However, interpreting these coefficients on plan types as direct evidence of an impact of network restrictions could be misleading if these network restrictions are correlated with other features of plan design such as cost-sharing.

[20] Contributors to Marketscan can be of to types: employers and health plans. The exclusion restriction $\mathbb{E}[\text{share}(p')_{et}\epsilon_{ispt}] = 0$ is more sensible for employers than for health plans.

Table 4. Wait Times, Patient Demographics and Insurance Plan Type

Dependent Variables:	Wait Time (days)		Wait Time (Pct SD)		Wait Time (days)		Wait Time (Pct SD)	
Model:	(OLS 1)	(OLS 2)	(OLS 3)	(OLS 4)	(IV 5)	(IV 6)	(IV 7)	(IV 8)
Panel A: Demographic Variables								
0-17	-9.672***	-9.869***	-8.265***	-8.372***	-9.700***	-9.883***	-8.285***	-8.384***
	(2.816)	(2.791)	(2.229)	(2.215)	(2.793)	(2.750)	(2.217)	(2.193)
18-34	0.7741	0.7783	-0.8395	-0.8372	0.8605	0.8662	-0.7936	-0.7905
	(2.080)	(2.087)	(1.183)	(1.188)	(2.100)	(2.109)	(1.193)	(1.200)
45-54	-1.367	-1.239	-0.0654	0.0040	-1.191	-1.008	0.0237	0.1221
	(1.143)	(1.150)	(0.6179)	(0.6247)	(1.125)	(1.136)	(0.6104)	(0.6201)
55-64	6.693***	7.189***	4.147***	4.416***	7.382***	8.021***	4.495***	4.838***
	(1.876)	(1.886)	(1.025)	(1.034)	(1.974)	(1.982)	(1.067)	(1.077)
Charlson Comor. Score	7.987***	7.981***	4.665***	4.661***	8.005***	7.998***	4.673***	4.669***
	(1.617)	(1.617)	(0.8559)	(0.8552)	(1.621)	(1.621)	(0.8573)	(0.8566)
Male	-21.30***	-21.53***	-10.83***	-10.96***	-21.35***	-21.59***	-10.86***	-10.99***
	(1.920)	(1.924)	(0.7164)	(0.7176)	(1.923)	(1.927)	(0.7125)	(0.7140)
Panel B: Insurance Design Variables								
HMO	5.288***	0.0668	2.651***	-0.1826	12.59***	7.177***	6.620***	3.714***
	(1.189)	(0.9843)	(0.6154)	(0.5705)	(2.277)	(1.979)	(1.199)	(1.106)
EPO	-0.6610	-7.609***	-0.2997	-4.070***	-8.201	-16.32**	-4.572	-8.931***
	(2.011)	(2.155)	(1.154)	(1.218)	(6.485)	(6.359)	(3.440)	(3.344)
POS	-4.514***	-7.836***	-2.447***	-4.250***	-7.352***	-7.375***	-4.126***	-4.138***
	(0.7590)	(0.9565)	(0.3768)	(0.4531)	(1.245)	(1.327)	(0.6784)	(0.7147)
POScap	4.626	-1.302	2.511	-0.7062	39.26***	13.68*	21.60***	7.861*
	(3.037)	(2.935)	(1.562)	(1.515)	(9.593)	(7.955)	(5.284)	(4.615)
CDHP	-4.240***	1.490	-2.626***	0.4837	-4.211*	5.299*	-3.109**	2.000
	(1.607)	(1.876)	(0.8354)	(0.9627)	(2.502)	(2.924)	(1.334)	(1.510)
HDHP	-11.97***	-1.440	-6.471***	-0.7533	-11.58**	1.792	-5.506**	1.678
	(1.807)	(1.713)	(0.8372)	(0.8545)	(5.079)	(4.930)	(2.682)	(2.647)
Comprehensive	-2.948	0.5924	-1.467	0.4547	-8.080**	-3.115	-3.870**	-1.203
	(1.940)	(1.805)	(0.9510)	(0.9001)	(3.447)	(3.215)	(1.695)	(1.605)
Share Out of Pocket		-52.05***				-64.48***		
		(4.347)				(5.941)		
Share Out of Pocket (Pct SD)				-0.0331***				-0.0406***
				(0.0021)				(0.0028)
Fit statistics								
Observations	786,318	786,318	786,318	786,318	786,318	786,318	786,318	786,318
R^2	0.22345	0.22419	0.00944	0.01038	0.22308	0.22386	0.00895	0.00998

Notes: Signif. Codes: ***: 0.01, **: 0.05, *: 0.1. Clustered surgery–year standard-errors in parentheses. Wait times with detection threshold at $\bar{p} = 0.9$. Instruments are average plan characteristics by contributor. All first-stage are strong as employers tend to offer few plans. All specifications include surgery–year, surgery–month fixed effects.

To separate the role of cost sharing from other aspects of plan design, we construct average out-of-pocket share (OOP) first at the patient and then at the plan level as follows:[21]

$$OOP_{ipt} = \frac{\sum_{c \in \mathcal{C}_{ipt}} \text{copay}_{cipt} + \text{deduct}_{cipt} + \text{coins}_{cipt}}{\sum_{c \in \mathcal{C}_{ipt}} \text{pay}_{cipt}} \text{ and } OOP_{pt} = \frac{\sum_{i \in p} OOP_{ipt}}{\sum_{i \in p} 1} \tag{5}$$

where $\mathcal{C}_{ipt}$ is the set of claims c in year t for patient i in plan p and co-payment, deductible and coinsurance are payments borne by the patient. The remainder is covered by the employer and the insurer. We investigate the impact of cost sharing on wait times within plan type by adding OOP_{pt} as an additional plan characteristic regressor $\boldsymbol{y}_{pt}$ in regression 4. In the same logic as with plan type variables, we address selection into plans within employer by instrumenting for out-of-pocket share at the plan level OOP_{pt} with the average out-of-pocket share at the employer level, $O\bar{O}P_{et}$. Figure C18 in Appendix shows the average out-of-pocket share for the main plan types between 2007 and 2013. Out-of-pocket shares vary starkly across plan types. HMO, EPO, and capitated POS plans have the lowest out-of-pocket shares on average, with patients bearing slightly less than 20% of all payments. For PPO and POS plans, between 25% and 30% of payments are covered by patients. Patients in high deductible plans (HDHP and CDHP) are responsible for between 40% to up to 50% of payments. This ranking in average out-of-pocket share by plan types matches the results from column (OLS1), (OLS3), (IV5) and (IV7) in Panel B of Table 4, which suggests longer wait times for HMO plans relative to PPO, which itself has longer wait times than CDHP and HDHP. We report the estimated coefficients from OLS and IV with the additional endogenous variable OOP_{pt} and the additional instrument $O\bar{O}P_{et}$ in levels and standard deviations in columns (2), (4), (6) and (8) in Panel B of Table 4. Even holding plan type fixed, the IV estimate implies that an increase in out-of-pocket share by 0.1 –about the difference between an average PPO and an average HMO– implies a drop in wait times by 6.4 days (column (6)). The inclusion of instrumented out-of-pocket share changes the magnitude of the estimated effects of plan types: the estimated effect of HMO is about 50% smaller in columns (6-8) than in columns (5-7), suggesting that half of it was due to difference in cost-sharing between HMO and PPO plans, and the rest can be attributed to other aspects of plan design, including network restrictions. The negative effect of high-deductible plans in columns (6-8) is no longer significant, suggesting that most of the difference in wait times between these plans and PPOs can be attributed to differences in cost-sharing. Figure C17b in Appendix displays a similar pattern for out-of-pocket share effects as what is observed on Figure C17a for the effect of HMO plans on wait times, with opposite signs. While out-of-pocket share reduces wait times across all surgeries except transplants, the estimated effects are larger for elective surgeries such as cataract and joint replacements.

To further outline the mechanisms at play, we study the effects of insurance plans on various com-

[21] Marketscan provides a plan key to identify plans, but in practice this plan key is only available for about 15% of patients in our sample. Instead we define a plan as the combination of an contributor (employer) and a plan type.

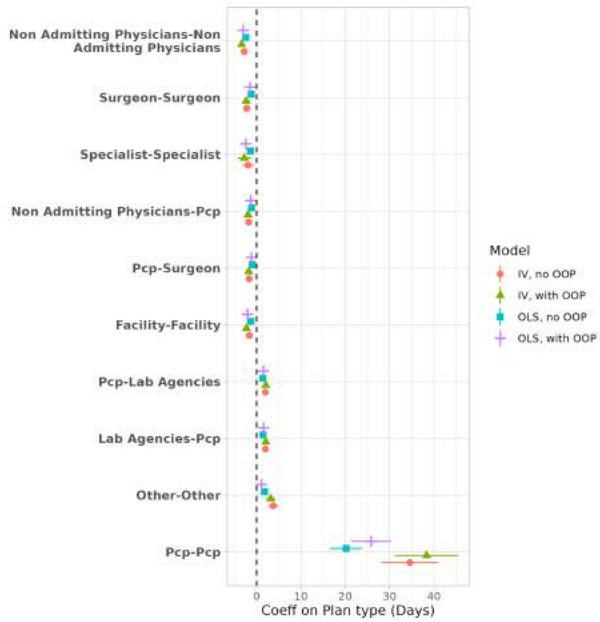
ponents of patients' trajectories that make up wait times. Marketscan contains detailed information on the type of medical professional encountered by patients at each visit. We group these detailed professions into 8 mutually exclusive categories, Surgeon, Specialist, Laboratories, Primary Care Physician, Non-Admitting Physician, Facility, Professionals, and Others, and we treat these categories in descending order of priority, so that if a patient sees a Surgeon and a Specialist during the same visit (day), the visit gets assigned Surgeon. Patients' wait times can then be broken down into 81 origin-destination pairs representing the time spent waiting between each successive categories as follows:

$$w_{ist} = \sum_{o,d \in \{\text{Surgeon},\ldots,\text{Other}\}} w_{is}^{od}$$

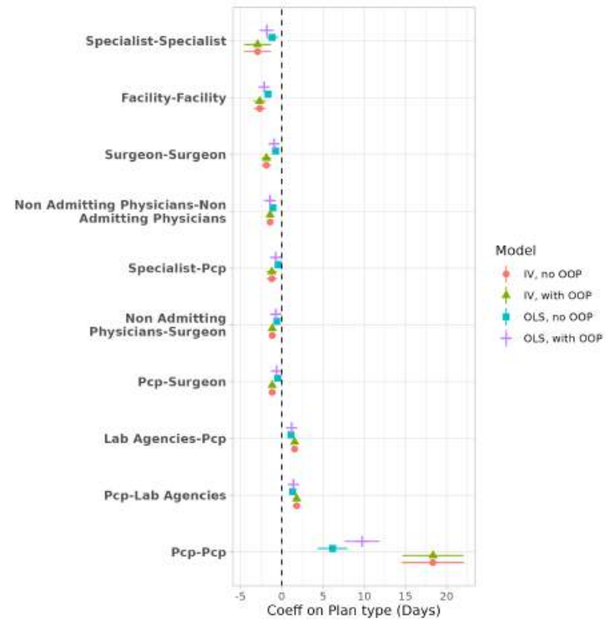
where a $w_{is}^{od} = 0$ if patient i has never had a visit with provider of type d following type o. Figure C19 and C20 in Appendix show the fraction of wait time spent on average in the 5 most important origin-destination pairs of categories by surgery. For the vast majority of surgeries, the first or second most important component of wait times is between 2 primary care physician (PCP) appointments. Time between two surgeon visits is also an important component, in particular for more emergency procedures such as coronary bypass, inguinal hernia or endarterectomy. To highlight how insurance design choices impact separate components of the wait times differently, we run the same specification as 4 with OLS and IV estimators but replacing the left-hand side variable with each of the origin-destination pairs w^{od}:

$$w_{ispt}^{od} = \boldsymbol{x}_i' \boldsymbol{\beta}_{\boldsymbol{x}}^{od} + \boldsymbol{y}_{pt}' \boldsymbol{\beta}_p^{od} + \phi_{st}^{od} + \epsilon_{ispt}^{od} \tag{6}$$

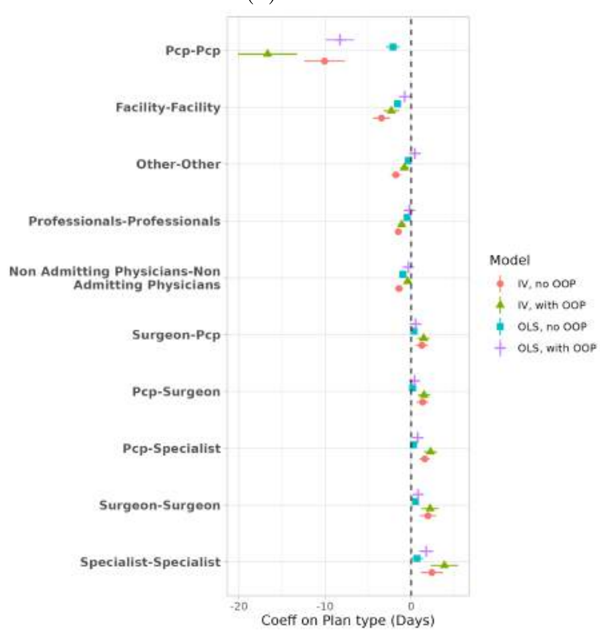
Figure 5 below shows the OLS and IV coefficients $\hat{\beta}_{p'}^{od}$ with largest absolute values for a subset of plans p', in specifications with and without plans OOP shares. The top panels show coefficients for HMO and POS, two managed-care plans relatively more restrictive in their design than PPOs. The main design difference between these plans and a PPO plan is the assignment of a PCP from which patient need to get referrals to see a specialist. We find that both plan types greatly increase time spent between PCP visits for patients, with an IV estimate of about 35 extra days for HMO and 18 days for POS, which is consistent with the PCP assignment preventing patients from choosing among generalists to reduce wait times, and also with previous studies showing that gate-keeping reduces specialist care in favor of generalist care (Sripa et al., 2019; Garrido et al., 2011). In contrast, patients in high-deductible plans tend to have lower wait times spent in between two PCP visits. Most of the important coefficients are negative, consistent with the overall shorter wait times associated with these plan designs in Table 4. Figure C21 in Appendix shows the coefficients of plan types on the number of visits by origin–destination pairs; these extensive margin effects go in the same direction as the estimates on wait times. Table C5 in Appendix shows in addition that patients in restrictive plans such as HMO or POS tend to go trough fewer visits during wait times (column (7)). However patients in HMOs wait on average 1.1 day more between two visits compared to patients in PPOs. The combination of fewer visits but longer wait between visits could be consistent with a plan design placing barriers for patients to proceed to the next stages of their care trajectory. Related to this idea, Table C6 shows that patients in HMO and POS plans receive more laboratory visits, but fewer preoperative tests than comparable PPO patients; this could suggest that patients in HMO and POS are required to multiply the types of providers they interact with before proceeding to surgery, without necessarily getting a more thorough preoperative testing. In addition, the last rows of Table C5 and C6 Appendix show that cost-sharing –within plan types– is associated with fewer visits, laboratory visits, and preoperative tests. This could be consistent with plans with higher cost-sharing also enforcing fewer restrictions on access to surgery. Alternatively, it could result from price-sensitive patients cutting back on care, a phenomenon evidenced when employers switch to high-deductible plans (Brot-Goldberg et al., 2017) or when Medicare patients bear a higher fraction of drugs' costs (Chandra et al., 2021).



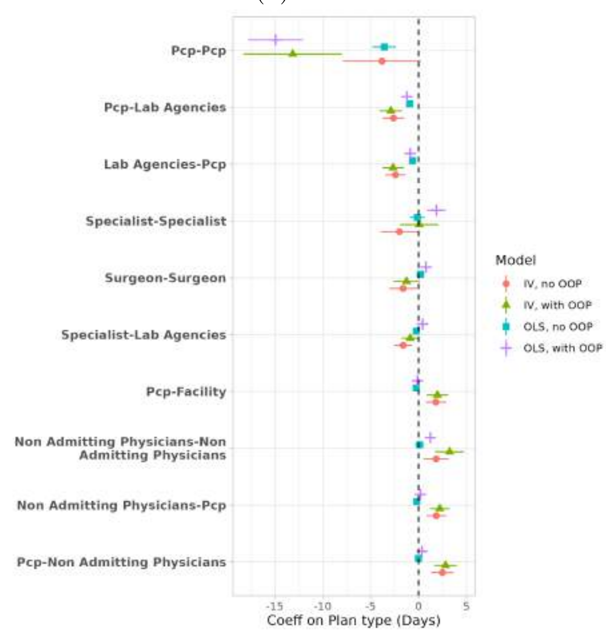
(a) HMO



(b) POS



(c) CDHP



(d) HDHP

Figure 5. Insurance Design and Components of Wait Times

Notes: Each Panel shows OLS and IV coefficients and 95% confidence intervals for an insurance plan type relative to PPO, estimated jointly. Specifications either include or exclude controls for plans' OOP share (instrumented in the case of IV). Sample excludes Healthplan contributors. Standard errors are clustered surgery-year .

5 Effects of Wait Times on Recovery post-Surgery

In this section, we study the impact of exogenous variation in wait times on patients' health. We rely on variation within surgery–insurance plan across weeks to isolate delays due to capacity constraints at the network level and plausibly unrelated to patient's unobserved health condition. This congestion design provides a strong first-stage across all our surgeries, highlighting yet another benefit of studying wait times using large-scale insurance claims data. We show that longer wait times have adverse consequences on a battery of health outcomes including inpatient re-admissions and spending, and prescriptions of opioids and addictive pharmaceutical drugs. We investigate how these adverse effects vary across surgeries and patient characteristics and use these heterogeneous treatment effect estimates to evaluate the medical efficiency of the allocation of wait times to surgery in the U.S. employer-sponsored insurance context.

Wait Times, Health Outcomes. To evaluate the impact of additional wait times on medical recovery, we look at the time period from the day after the surgery to six months after and compute the following medical outcomes : (1) **inpatient payments** defined as the sum of all payments in Inpatient Services table, (2) **outpatient payments** defined as the sum of all payments in Outpatient Services table, (3) **inpatient length of stay** defined as the sum over all inpatient admissions episodes of the difference between discharge and admission dates, (4) **inpatient readmission** defined as an indicator for any readmission over the period, (5) **drug payments** defined as the sum of all payments in Outpatient Pharmaceutical Drugs table, (6) **days supplied of opioids** defined as the sum of days supplied of generic drugs from the *opiate agonist* class and (7) **days supplied of addictive drugs** defined as the sum of days supplied of generic drugs from Class II (high abuse potential, severe dependence liability) in the Drug Enforcement Administration's classification of controlled substances. Denoting these outcomes as y, we run the following regressions:[22]

$$y_{ispet} = \beta_0^{(y)} + \beta_w^{(y)} w_{ispet} + \boldsymbol{x}_i' \boldsymbol{\beta}_x^{(ys)} + \phi_{st}^{(y)} + \psi_{spe}^{(y)} + \epsilon_{ispet}^{(y)} \tag{7}$$

Identification Challenges. Patient controls $\boldsymbol{x}_i$ include gender, age group, and Charlson Index, a measure of a patient's health status based on the presence of severe diagnoses.[23] In addition, the specification includes surgery–year, surgery–month, and surgery–insurance plan fixed effects. However, even with these controls, doctors might allocate patients to care faster based on unobserved cues about patients' health picked up during face to face visits which would create a negative correlation between wait times w_{ispet} and the error term $\epsilon_{ispet}^{(y)}$ and could subsequently generate attenuation bias in the coefficient of interest $\beta_w^{(y)}$.

[22] We report the results with wait times in levels in the main text, but we find effects of similar magnitudes with wait times in log. We also intend to implement Poisson regressions as an alternative model.

[23] Later in our exercises with double machine learning, we include hundreds of controls measuring the type and quantity of medical services a patient consumes. We recover even higher average treatment effects.

Congestion Design. To address these concerns, we propose an instrumental variable strategy relying on congestion within a patient's insurance network. In a first step, we construct shocks in delays that shift a patients wait time. For each patient i, wait time is composed of multiple segments between two successive visits. These visits involve various provider types; we aggregate provider types recorded in Marketscan into 8 mutually exclusive categories, Surgeon, Specialist, Laboratories, Primary Care Physician, Non-Admitting Physician, Facility, Professionals, and Others. For each patient i we identify the longest segment among the sub-components of wait time, the corresponding provider type $k^*(i)$ patient i visits at the end of the segment, and the date $d^*(i)$ at which that segment ends. We construct a congestion shock g_{ipe} to i's longest waiting segment by computing the average wait time w for provider type $k^*(i)$ at date $d^*(i)$ for all other patients in i's contributor e–plan p type:[24]

$$g_{ipe} = \frac{\sum_{j \neq i, j \in pe} w_{jd^*(i)k^*(i)}}{\sum_{j \neq i, j \in pe} \mathbf{1}_{jd^*(i)k^*(i)}}$$

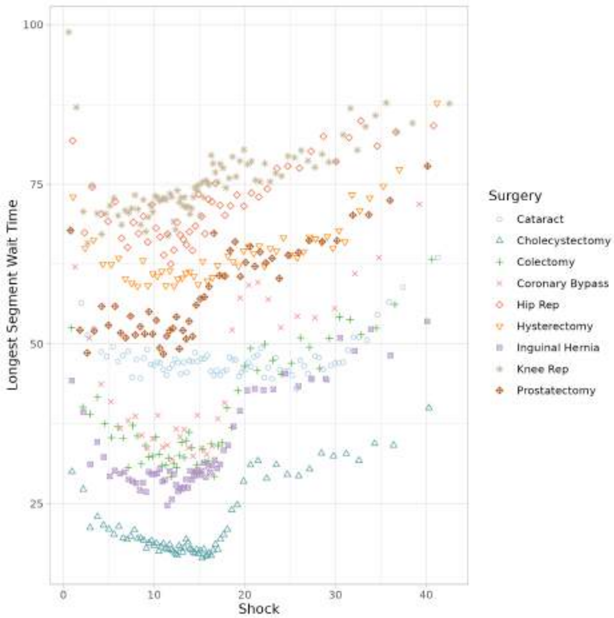
The assumption underlying the design is that the shock g_{ipe}, which only involves the wait times experienced by patients others than i, is orthogonal to unobserved drivers of patient i's health. Figures 6c and 6d show the distributions of these shocks by provider types. Figure C25c below illustrates the congestion effect of the shock on a patients longest segment. Across surgeries, for large enough values of the shock, wait time experienced on patients longest segment increases; however, this congestion effect is not present at lower values of the shock. In particular, in the range of shocks where most of the mass is located, as can be seen on the histograms in figures 6c and 6d, the effect of the shock on wait time at that leg is clearly positive. This non-linearity motivates a second-step in constructing the instrument in which we compute the best predictor of wait time $z_{ispe} = \mathbb{E}[w|g_{ipe}, s]$ conditional on the longest segment congestion shock g and surgery s.[25] Figure 6b shows an unconditional bin-scatter plot representing the fit of this predicted wait time by surgery. The instrument leverages variation in wait times induced by the limited capacity of a patients networks of providers. Additional wait times could stem from at least two causes: direct supply shocks, such as a doctor unexpectedly cancelling appointments for personal reasons, or demand shocks *from other patients* leading to delays induced by congestion. A first concern is that insurance plans and their resulting networks are chosen by patients, which could lead to a selection problem. We address this with insurance plan by surgery fixed effects, so that the variation we rely on is limited to within insurance plan changes in congestion. Similarly, seasonal demand shocks are absorbed by surgery by month fixed effects. Another potential issue is within network selection of provider types. In the next section of the analysis investigating heterogeneous effects, the use of machine learning based estimators enable us to include as controls the entirety of patients medical history from their claims, in particular the counts of visits at all potential provider types. If anything, the average effects that we recover with these high-dimensional controls are higher than

[24] In practice, to get enough patients in the cells used to construct congestion shocks, we take d^* to be an entire week instead of just a single date. Moreover, we exclude shocks that rely on fewer than 50 patients.

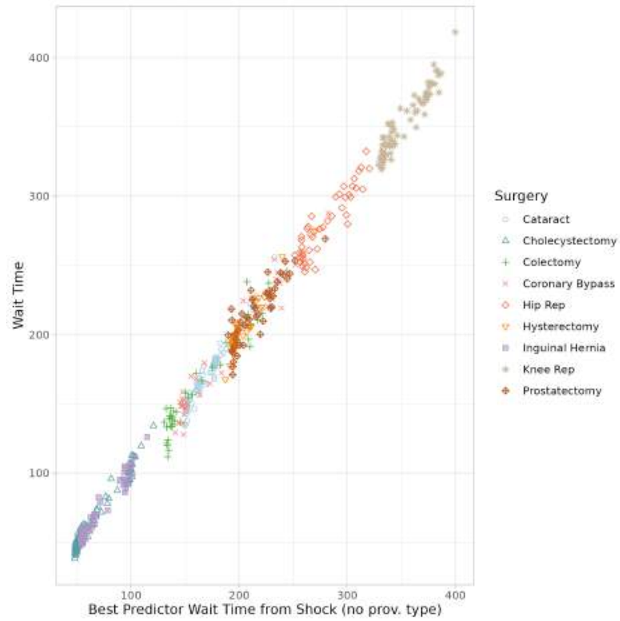
[25] We use the out-of-bag predictions of a random forest to compute this best predictor.

the ones obtained with the more parsimonious linear-IV estimators. A related concern is selection of providers within networks and provider types: patients with more serious health conditions could seek higher quality doctors who are also more likely to be over-subscribed. However, the instrument that we use relies on variation in congestion *averaged* over all providers in a given type. The most frequent provider types at the origin of the congestion shocks are PCP and specialists; within a given plan, there will typically be many providers belonging to each of these types, which implies that the congestion shock is averaged across a large number of providers in a given week. This averaging alleviates the concern that selection on provider quality could lead to selection on congestion. Moreover, we run the linear-IV regressions on a restricted sample of patients enrolled in managed-care plans. By design, these plans limit the flexibility with which patients can select the specific doctor they want to see. Therefore, if the selection of sicker patients to congested providers was creating exaggeration in the average effects reported in our main results, we would expect estimated effects to be smaller in the managed-care sample. As can be seen when comparing appendix Tables C10 and C11 for the managed care sample to appendix Tables C8 and C9 for the main sample, this is not the case: the estimated effects are very similar in both samples. Finally, appendix Table C7 reports balance for age and Charlson index between patients in the fourth and first quartiles of wait times and of the instrument.[26] As anticipated, the instrument displays much better balance than wait times, which we showed in the previous section tend to be longer for women, older, and patients with more chronic conditions. While for wait times, standardized mean differences of more than 20 percent of pooled standard deviation in absolute value are common, for the instrument for the majority of surgeries the difference is below 6 percent in absolute value, and many surgeries have mean differences below 2 percent.

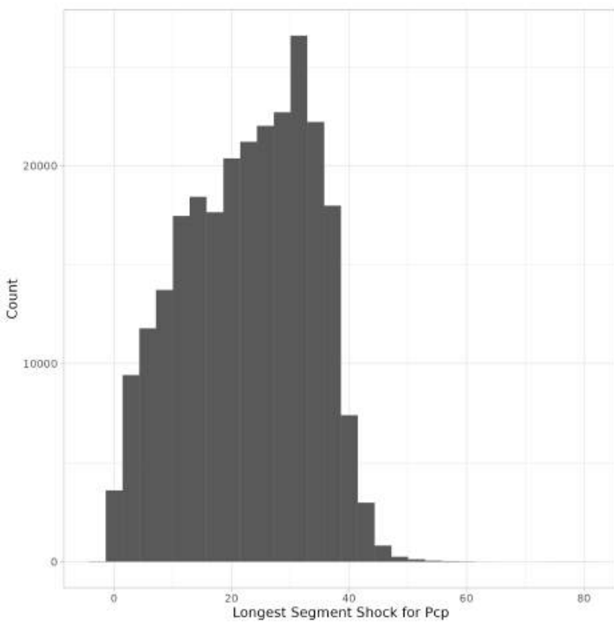
[26] We do not need to assume that congestion is orthogonal to observed patient characteristics, and in fact control for age and Charlson index in our specification.



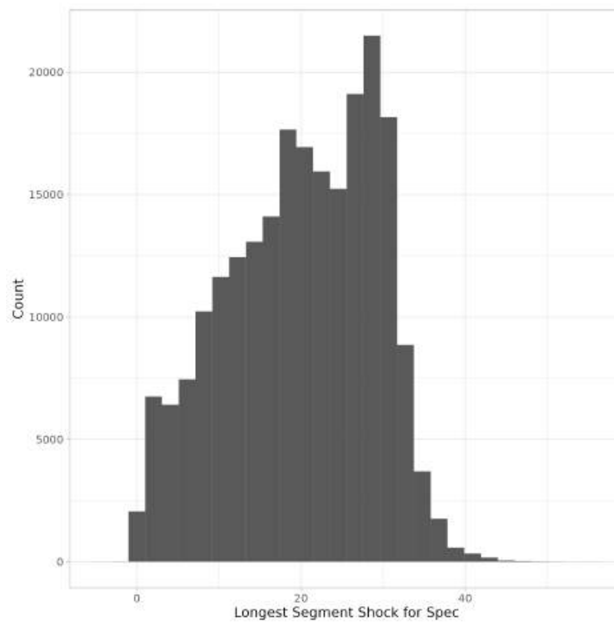
(a) Delay at Segment



(b) Predicted Wait Time



(c) Congestion PCP



(d) Congestion Specialist

Figure 6. Variation in Wait Times from Congestion Shocks

Notes: Panels a and b show binscatter plots. On panel b, the y-axis the out-of-bag prediction of wait time from a random forest given only the congestion shock and the surgery a patient is waiting for.

Table 5. Impact of Wait Times on Health Outcomes

Dependent Variables:	Inp. Pay ($)	Outp. Pay ($)	Inp. Stay (days)	Inp. Readmission (pct.)	Inp. Pay ($)	Outp. Pay ($)	Inp. Stay (days)	Inp. Readmission (pct.)
			OLS				IV	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Wait Time 0.9 (days)	2.047***	-0.2681***	0.0005***	0.0041***	7.251***	-0.9909	0.0014***	0.0134***
	(0.1793)	(0.0900)	(3.08×10^{-5})	(0.0002)	(1.751)	(0.8018)	(0.0003)	(0.0018)
Effect +1 Month Wait (pct.)	1.7	-0.1	1.8	1	5.9	-0.4	5.1	3.1
Outcome Mean	3667	6860	0.79	12.77	3667	6860	0.79	12.77
Fixed-effects								
Year, Month and Plan F.E.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics								
Observations	762,466	762,466	762,466	762,466	762,466	762,466	762,466	762,466
R^2	0.06173	0.32510	0.07811	0.10167	0.06019	0.32503	0.07668	0.09917
F-test (1st stage), Wait Time 0.9 (days)					12,609.6	12,609.6	12,609.6	12,609.6

Heteroskedasticity-robust standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Wait times with detection threshold at $\bar{p} = 0.9$. Sample excludes Appendectomy, Nephrostomy, and Splenectomy for which the detection models perform poorly, and only includes patients from employer contributors. Dollars are deflated to 2010 dollars.

Table 6. Impact of Wait Times on Addictive Drugs Consumption

Dependent Variables:	Drug Pay ($)	Opioids (days supp.)	Addict. Drugs (days supp.)	Drug Pay ($)	Opioids (days supp.)	Addict. Drugs (days supp.)
		OLS			IV	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Wait Time 0.9 (days)	0.7601***	0.0135***	0.0079***	1.840***	0.0397***	0.0209***
	(0.0217)	(0.0003)	(0.0003)	(0.1671)	(0.0025)	(0.0020)
Effect +1 Month Wait (pct.)	1.8	2.2	2.6	4.3	6.6	7
Outcome Mean	1291.9	18.13	9.01	1291.9	18.13	9.01
Fixed-effects						
Year, Month and Plan F.E.	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	762,466	762,466	762,466	762,466	762,466	762,466
R^2	0.13971	0.10552	0.04904	0.13623	0.09593	0.04520
F-test (1st stage), Wait Time 0.9 (days)				12,609.6	12,609.6	12,609.6

Heteroskedasticity-robust standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Wait times with detection threshold at $\bar{p} = 0.9$. Sample excludes Appendectomy, Nephrostomy, and Splenectomy for which the detection models perform poorly, and only includes patients from employer contributors. Dollars are deflated to 2010 dollars.

Results. We present the estimates from OLS and IV estimators for inpatient and outpatient-based outcomes in Table 5 and for drugs-based outcomes in Table 6.[27] First, comparing coefficients across estimators reveals that the ones estimated by OLS are smaller than the IV ones, suggesting that prioritization of unobservably sicker patients by the health care system might indeed be at play. Tables 5 and 6 report the estimated coefficients from a specification in levels of wait time. The tables also report the implied causal impact of an additional month of wait times on outcomes in percentage terms. We find that delays in access to surgery negatively impact patients' recovery as measured by multiple adverse medical outcomes: the estimated coefficients are positive and significant for all outcome variables except outpatient payments. The estimated effects are large: an extra month of wait times—corresponding to an 18% increase, or about the average difference in wait times between a man in a PPO plan and a woman in an HMO plan implied by the analysis in Section 4—would translate into an increase in post-surgery inpatient payments of 5.9%, inpatient length of stay longer by 5.1%, inpatient readmissions by 3.1%, an increase in opioid intake of about 6.6% and 7% more days supplied of addictive drugs. Figure C22 in appendix shows the causal effect of an additional month of wait time implied by instrumental variable regression results, but this time split by outcome and by surgery. First, with the exception of outpatient payments for which the average effect is negative, insignificant, all outcomes display a similar pattern where the average effects are either null or positive. This is consistent with the intuition that exogenous increases in wait times should be either detrimental or have no detectable effect on health outcomes, but cannot improve patient health. For outpatient payments, the negative estimates for colectomy, mastectomy and thyroidectomy could imply that patients who wait longer substitute away from outpatient care towards hospital settings.

Persistence. We study how long-lasting the adverse effects associated with delays are by running regression 7 while varying the window of time in which medical outcomes are computed from a month after the surgery to between five and six months after the surgery. Figure C23 in appendix shows the coefficients from IV estimators for each outcome for which we find a statistically significant effect and for each month after the surgery. For outcomes based on inpatient visits, the coefficients show that most of the effect is driven by the first month after the surgery, pointing to complications that require readmission directly following the initial surgery. The estimated IV coefficients are around 2.5 in the window of time ranging from 1 to 30 days from the surgery –about a third of the cumulative effect over the whole period– as opposed to 0.75 between 150 and 180 days from the surgery –or about a tenth of the overall cumulative effect. In spite of this decline, the estimated effects of additional wait times are still positive and significant for virtually all inpatient outcomes even six months after the surgery. For opioid and addictive drugs intake represented in Figure C23e and C23f, there is very little decline, which suggests that delays cause a constant and persistent increase in addictive drugs prescriptions and in particular in opiate agonists. There

[27] In appendix Tables C8 and C9 we present results for inpatient payments and prescribed opioids only, this time varying specifications. The last column in these tables correspond to our preferred specification shown in the main text.

could be at least two mechanisms generating this result: one is a consequence of preoperative prescriptions, whereby patients who get prescribed opioids and end up waiting longer develop a dependency which makes it more difficult to fade out the opioid intake post-surgery. The second is a direct channel in which wait times decrease the success of surgeries, leaving patients in pain and therefore requiring more opioid prescriptions for an extended period of time after the surgery.

Heterogeneous Effects. While these results on the average impact of wait times on health outcomes shed light on the medical costs associated with regulating access to care via queues, we expect delays to be especially detrimental for some subgroups of patients waiting for specific procedures. Estimating heterogeneous effects is a crucial step towards designing better wait-lists: identifying the subgroups who suffer the most from longer wait times would enable insurers to prioritize these patients. Doing so in this setting presents several challenges. First, identifying heterogeneous effects typically requires observing many surgical patients to have enough power. Moreover, in the case of the impact of wait times, endogeneity concerns call for a careful empirical design. Finally, the dimensions across which patients differ are large, with many possible medical histories leading to the same surgery. The methodology developed in this paper allows us to address these challenges. First, the measure of wait time we develop can readily be computed for thousands of patients at once. Second, we propose a congestion design that generates an instrument with a strong first stage for every surgery included in our study. With this design, and the rich medical controls provided by claims data, we implement a double machine learning pipeline to estimate heterogeneous marginal effects. We outline our approach below.

Model Selection and Validation. Machine learning for the estimation of heterogeneous effects typically consists in learning nuisance functions in a first step, in our case the conditional expectations of outcomes, treatments, and instruments, then residualize these variables using cross-fitting, and finally learn heterogeneous effects by minimizing a moment condition constructed from these residualized variables.[28] In the case of continuous treatments and instruments, several approaches have been developed, the most frequently used being generalized random forests (GRF) (Athey et al., 2019) and double machine learning with instruments (DMLIV) (Chernozhukov et al., 2018a,b; Syrgkanis et al., 2019).

With these methods, we estimate the following model:

$$\begin{aligned} y &= \theta(X).w + f(X) + \epsilon \\ \mathbb{E}[\epsilon|z, X] &= 0 \end{aligned} \tag{8}$$

where y is our medical outcome of interest, X is a large-dimensional matrix including fixed effects,

[28] The specific moment condition depends on the algorithm.

controls for patients demographics and medical history, w is wait time, z is the instrument based on congestion described in the previous section. The estimand of interest is $\theta(.)$, the heterogeneous causal effects of wait time w on outcome y as function of patient characteristics X. Finally, note that model 8 is restrictive in that it assumes a linear effect of wait time on outcomes. In principle we could allow for richer models of the form $\boldsymbol{\theta}(X).\boldsymbol{\phi}(w)$ where $\boldsymbol{\phi}(.)$ would contain non-linear transformations of wait times such as higher order terms.[29] We start with a linear effect model but also experiment with quadratic $\begin{bmatrix}\theta_1(X)\\ \theta_2(X)\end{bmatrix}.\begin{bmatrix}w\\ w^2\end{bmatrix}$ and log effect models. As is typical in this literature, we train several models and evaluate them out-of-sample.[30] Relative to model selection for standard predictions problems, the estimation of treatment effects presents the additional challenge that counterfactual outcomes for an individual are never observed, referred to as the fundamental problem of causal inference (Holland, 1986). While model selection for heterogeneous effects estimation has been studied recently (Mahajan et al., 2022; Doutreligne and Varoquaux, 2023), these papers tend to focus on models where conditional unconfoundedness is assumed. To guide our model selection, we use a calibration approach such as the one used in Athey et al. (2023), and adapt it to a setting where the estimation requires the use of instruments. We define the calibration score used for model selection below. First, we train nuisance functions $\mu(X) = \mathbb{E}[y|X]$, $\pi(X) = \mathbb{E}[w|X]$, $\rho(X) = \mathbb{E}[z|X]$ in a cross-fitting way.[31] These functions allow us to compute the double machine learning estimator of average treatment effect with instrument (DMLATEIV) (Syrgkanis et al., 2019):

$$\tilde{\theta} = \frac{\mathbb{E}_X[(y - \hat{\mu}(X))(z - \hat{\rho}(X))]}{\mathbb{E}_X[(w - \hat{\pi}(X))(z - \hat{\rho}(X))]} \tag{9}$$

Second, we train our heterogeneous effects models using 5-fold cross validation: a model trained on 80% of the data is used to predict heterogeneous effects $\hat{\theta}(X^{test})$ in the remaining 20% of the data. We use these predicted effects to split the test-set into quantiles of treatment effects, from most to least impacted by wait times. In each of the quantiles $g \in 1, ..., G$, we compute the average treatment effect implied by our heterogeneous effects model $\hat{\theta}_g = \frac{\sum_{i \in g} \hat{\theta}(X_i)}{\sum_{i \in g} 1}$ and estimate the average treatment effect using DMLATEIV, $\tilde{\theta}_g$. Note that while $\hat{\theta}_g$ is the average of predictions from the trained function $\hat{\theta}(.)$ in quantile g and therefore only uses the matrix of controls X_g, $\tilde{\theta}_g$ estimates the average effect using outcomes, wait time, instrument and controls y_g, w_g, z_g, X_g. If the two quantities $\tilde{\theta}_g$ and $\hat{\theta}_g$ coincide, it would not only suggest that the estimated heterogeneous effects function $\hat{\theta}(.)$ has successfully ranked out-of-sample patients according to the severity of their adverse consequences from waiting, but also that the magnitude of the predicted effects is in the

[29] There exists more flexible methods such as Hartford et al. (2017) which do not restrict the form of the response function. We choose to experiment with double machine learning and generalized random forests first on account of their relative simplicity.

[30] We use the python library EconML (Battocchi et al., 2019) to implement various models.

[31] We use histogram gradient boosting for each of these nuisance functions.

correct range. We define our calibration score with instrument, $C^{IV} \in]-\infty, 1]$ as follows:

$$C^{IV} = 1 - \frac{\sum_g |\tilde{\theta}_g - \hat{\theta}_g|}{\sum_g |\tilde{\theta}_g - \tilde{\theta}|} \tag{10}$$

The higher this score, the better the heterogeneous effects model is at predicting *out-of-sample* whether patients will be impacted by delays to access surgical care. A calibration score of 1 implies that the average effects estimated in the groups formed out-of-sample by the heterogeneous effects model coincide exactly with the ones estimated using DMLATEIV. Conversely, a negative score indicates that heterogeneous effects model is doing poorly at detecting most impacted groups out-of-sample, and that in fact using just the full sample average treatment effect $\tilde{\theta}$ would yield better predictions.

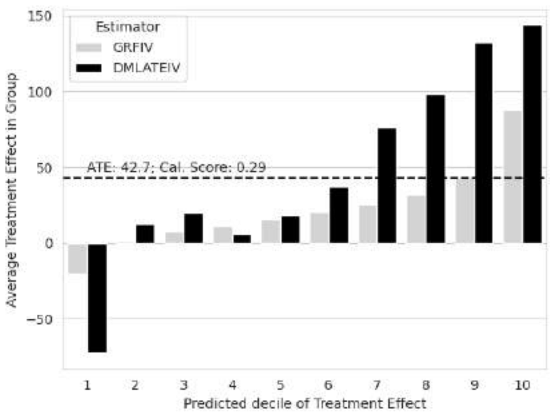
Results. Figure 7 displays the heterogeneous effects obtained with the best model –generalized random forests– for coronary bypass surgery. The top left panel for Figure 7 shows that (1) there is heterogeneity in the effects of wait time, as the high calibration score of about 0.3 suggests that the model outperforms a constant effect model and (2) that the heterogeneous effect model successfully predicts on patients never seen before the severity of their adverse effects of wait time based on patients medical trajectory and demographics. In fact, the average effects by quantile estimated using DMLATEIV are almost monotonically increasing. The heterogeneous effect model seems to slightly underestimate the severity of the impact of wait times for patients classified as at-risk, as can be seen in deciles 7 to 10 where the average effect predicted by GRFIV falls below the average effect estimated with DMLATEIV. This implies that the predicted effects implied by the GRFIV model can be thought of as lower bounds. The top right panel of Figure 7 shows the distribution of predicted effects in the population of patients waiting for a coronary bypass procedure. First, most of the distribution lies above zero, consistent with the intuition that longer wait times should not improve patient health. Second, this figure highlights the extent of heterogeneity in treatment effects: more than 10% of patients have predicted effects of more than $75 per extra day of wait time, or about 3 times the estimated average effect. Conversely, for about half of the of patients, waiting one more day is completely benign. The bottom left panel of Figure 7 plots the average wait time by decile of predicted effects, where the first decile are the patients the least impacted by waiting and the tenth decile are the patients the most impacted by waiting. Strikingly, this figure presents a increasing relationship between wait time and adverse effects of wait time, at the exact opposite of what a medically efficient allocation would require.[32] We note that this pattern is not present for other surgeries in our sample: the same figures for colectomy and knee replacement surgery in appendix (Figures C24 and C25) do not present a similarly increasing relationship, although none of these surgeries displays an allocation close to efficient either. Finally, the bottom right panel of

[32] One concern could be that this apparent inefficiency reflects a non-linearity in the effect of wait time. We experiment with quadratic effects model but find that they perform much worse in terms of predicting marginal effects out-of-sample.

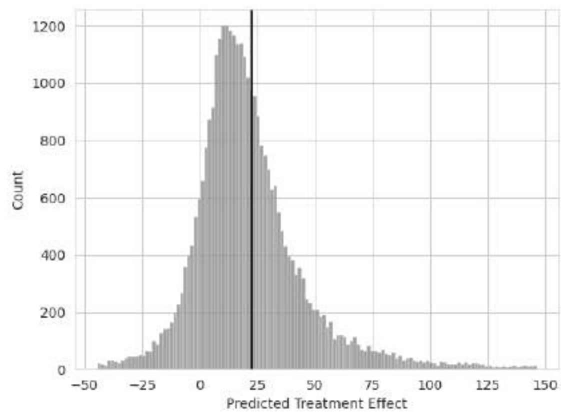
Figure 7 compares the characteristics –both demographics and medical– of patients most impacted by delays relative to the least impacted ones for coronary bypass surgery. The bottom right panels of Figures C24 and C25 show the same comparison for colectomy and knee replacement. For these three procedures, similar patterns emerge: patients adversely impacted by delays tend to have higher Charlson comorbidity score than patients with smaller estimated effects, in line with the medical intuition that patients with chronic conditions should be prioritized. Accordingly, more impacted patients consume more healthcare services, in particular tests such as electro-cardiograms (EKG) and blood counts, and are also more likely to consume inpatient care, which is associated with more serious health issues. Together, these profiles of at-risk patients suggest that our models identify credible dimensions of heterogeneity that matter for the extent to which delays will cause adverse outcomes for a particular patient. Finally, Figure C24 for colectomy shows an interesting pattern where a specific diagnosis, *Diverticula of the Intestine*, is strongly associated with lower adverse effects from wait times. This suggests that this cause for colectomy is not as pressing as say, *Malignant Neoplasm of Rectum* which we find is much more prevalent in patients severely impacted by wait times. We interpret this result as an additional validation of our heterogeneous effects model.

Mechanisms behind Misallocation. While striking, the extent of misallocation of surgical wait times we uncover is not surprising. First, wait times for elective surgery in the United States are not managed through a centralized mechanism that could help achieving certain desired medical objectives. As such, there are no explicit prioritization policies in place. Instead, the surgical wait times we measure result from repeated interactions of patients with multiple providers types, from PCP to surgeons to hospitals, where each step might be subject to specific supply constraints. A doctor at a given step might not have access to the full picture of a patient' health situation, making prioritizing decisions difficult. Moreover, the finding that patients who suffer the most from delays – typically older patients with a higher Charlson index reflecting comorbidities such as diabetes – wait longer is consistent with healthcare practices leading these patients to receive an unnecessary amount of medical tests in the preoperative period. These tests could follow from extra precautions providers might take for certain chronic conditions, such as stabilizing patient glycemia in a desired range before surgery. Dossett et al. (2022) highlight over-testing as a potential cause of surgical delays, and medical research finds that preoperative testing leads to delays for multiple surgeries such as heart surgery, (Coffman et al., 2019), cataract (Chen et al., 2021), and hip fracture repair (Ricci et al., 2007; Bernstein et al., 2016).

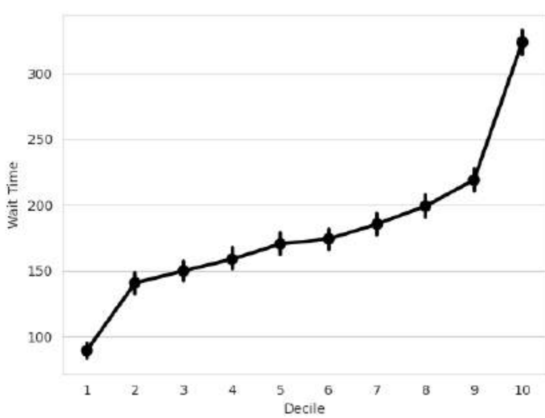
Quantifying Spending Reductions from Improved Allocative Efficiency. Our results highlight the presence of substantial heterogeneity in adverse effects of wait times. We also find that patients who suffer the most from delays are not necessarily the ones that wait the least, far from it. This medical misallocation implies potentially large efficiency gains in terms of hospital spending



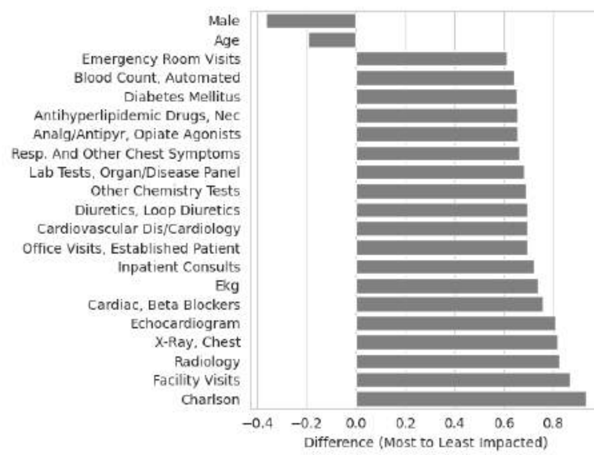
(a) Calibration Plot



(b) Distribution of Effects



(c) Allocative (In)Efficiency



(d) Characteristics of Most Impacted Patients

Figure 7. Heterogeneous Effects from Wait Times on Inpatient Spending for Coronary Bypass

Notes: Panel (a) displays the figure resulting from the calibration procedure in which we rank patients according to deciles of predicted effects from a heterogeneous effects model and then compare the average of the effects ***predicted*** by that model in each decile to the average effect ***estimated*** in each decile by DMLATEIV. The implied calibration score and the full sample average effect estimated by DMLATEIV are also reported on panel (a). Panel (b) shows the distribution and mean of heterogeneous effects. Panel (c) displays average wait times and standard deviations by decile of predicted effects, where the tenth decile contains the patients the most impacted by wait times. Panel (d) shows the difference in standardized patient characteristics between patients in the fourth quartile of effects and patients in the first quartile of effects.

from reducing wait times for targeted subgroups of patients. The fact that, for each surgery for which we can confidently compute heterogeneous effects, the majority of patients experience small or even no adverse effects from delays suggests that these spending reductions could be achieved without requiring a decrease in aggregate wait times, which we would interpret as an expansion of aggregate provider capacity. We quantify these gains through simulations exercises where wait times are reallocated across patients so as to minimize hospital spending subject to fixed health care sector capacity. To compute counterfactual spending, we first estimate expected spending based on patient covariates and wait times:[33]

$$\rho(w, X) = \mathbb{E}[y|w, X]$$

Given an estimate $\hat{\rho}(.)$, and our heterogeneous effects estimates $\hat{\theta}(.)$, the predicted spending $\hat{y}(w^{cf}, X)$ given covariates X and counterfactual wait times w^{cf} are given by:

$$\hat{y}(w^{cf}, X) = \hat{\rho}(w, X) + \hat{\theta}(X)[w^{cf} - w] \tag{11}$$

The optimization problem corresponding with the wait times reallocation exercise described above can be written as:

$$S^{cf} = \arg \min_{\boldsymbol{x} \in \{0,1\}^{I \times J}} \sum_{i,j} \hat{y}(x_{ij} w_j, X_i) \text{ s.t } \sum_j x_{ij} = 1 \text{ and } \sum_i x_{ij} = 1 \tag{12}$$

whereby patient i is assigned the wait time of a unique other patient j, so that her counterfactual wait time w_i^{cf} is given by $x_{ij} w_j$.[34] The solution $\boldsymbol{x}^*$ simply involves reallocating the shortest wait times times to the patients with the largest adverse effects from waiting $\hat{\theta}(X_i)$.[35] We denote this counterfactual as the *Optimal* scenario, reflecting the maximum extent of hospital spending reduction that can be achieved by reshuffling wait times across patients. We then compare the average per patient hospital spending in the 6 months following the surgery resulting from that *Optimal* scenario to the one predicted by our model with the *Observed* wait times allocation in the data, and to a *Random* allocation of wait times. Finally, we compute hospital spending in two additional counterfactuals: one where we constrain the reallocation to take place within insurance plan, called *Optimal by Plan*, and one where the reallocation is based on a coarser version of $\hat{y}$ in which we replace $\theta(\hat{X}_i)$ with $\mathbb{E}[\hat{\theta}(X)|X^{rule} = X_i^{rule}]$ where $X^{rule} = \{age, gender, Charlson\}$. This last counterfactual, called *Demographic Target*, captures the potential gains from designing prioritization policies based on simple rules using only a reduced set of covariates.

[33] We estimate a flexible $\rho()$ for each surgery using the automatic machine learning library autoML.

[34] The choice of an assignment problem treats not only total wait times, but also each individual trajectory as fixed. It is conservative in the sense that it would not require any change to the health care system, and only implies substituting patients with one another. A downside is that reallocation can involve large changes in wait times, implying that we are extrapolating effects away from the wait times at which they were estimated. As an alternative, we could instead experiment with linear programming problems where we only hold total wait times fixed, and individual wait times are only allowed to change within a specified percentage range.

[35] In the implementation, we winsorize the left tail of $\hat{\theta}(.)$ so that predicted negative effects are set to 0. In doing so, we prevent spurious counterfactual savings stemming from negative effects patients being allocated large wait times.

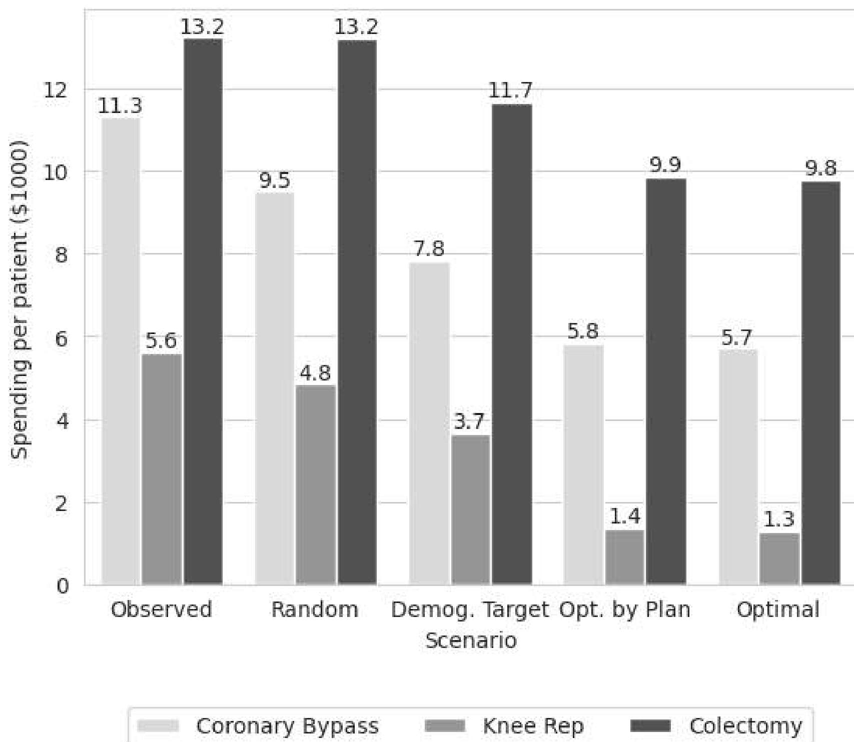


Figure 8. Spending Reduction from Efficient Wait Times Assignment

Notes: This figure shows the average per patient spending resulting from reallocating wait times across patients according to the constrained assignment problems described in the main text. Dollars are in 2010 USD. *Observed* corresponds to predicted hospital spending from observed wait times and covariates. In all of the other scenarios, wait times are reshuffled, keeping total wait times constant, and implied spending change according to estimated marginal effects. *Random* reallocates wait times at random. *Demog. Target* reallocates wait times according to average marginal effects by age, gender, and Charlson index. *Optimal by plan* reallocates within plan according to marginal effects. *Optimal* reallocates according to marginal effects without restrictions.

Gains from Improving the Efficiency of Surgical Wait Lists. Figure 8 displays the average hospital spending per patient in each of the counterfactual scenario described above. First, comparing the *Observed* and *Random* scenarios highlights the misallocation of surgical wait times which was already apparent in Figures 7 and C24:C25 the per patient spending falls when moving from observed wait times to a random allocation of wait times. Second, for all three surgeries, the reallocation scenarios *Demographic Target*, *Optimal by Plan*, and *Optimal* all greatly decrease hospital spending per patient. These aggregate hospital spending gains could mask heterogeneity; however, given the skewed distribution of marginal effects recovered in Figures 7 and C24:C25, we anticipate that patients with large adverse effects gain the most from seeing their wait time reduced, while patients who experience a wait time increase do so at little cost for their health. Comparing the gains between the *Demographic Target* and *Optimal* scenarios is instructive: while the gains are almost twice as large in the optimal scenario, this amount of targeting might not be feasible in practice. It is therefore encouraging to see that even with a simple targeting rule using only 3 patient characteristics as in *Demographic Target*, we are able to achieve large spending reductions. This shows that models derived from our measure and design could provide prioritization rules that (1) are shown in a causal setting to increase efficiency and (2) are no more complex than alternative rules already implemented in other settings such as Australia or Canada's joint replacements wait lists (Siciliani et al., 2013).

6 Conclusion

This paper develops a novel method to measure wait times for surgical care, relying only on insurance claims data. We implement this method in one of the most extensive claims datasets of U.S. patients with employer-sponsored insurance, a setting where administrative data on wait times are typically not observed. Our method could be applied to compute wait times in other contexts where administrative data is not systematically available. It could also serve as a basis for cross-country analyses or studies of how wait times have changed as populations age. Combining our measure with an empirical design based on within-network congestion, we show that wait times have adverse consequences on health that are very heterogeneous across and within surgeries. We use these estimates to quantify misallocation in surgical wait times for three surgeries and find large inefficiencies. With geographic or more detailed demographic data, our measure could shed light on a key dimension of health care equity: that of timely access to surgical care.

References

Agarwal, N., I. Ashlagi, E. Azevedo, C. R. Featherstone, and Ö. Karaduman (2019): "Market failure in kidney exchange," *American Economic Review*, 109, 4026–4070.

Agarwal, N., I. Ashlagi, M. A. Rees, P. Somaini, and D. Waldinger (2021): "Equilibrium allocations under alternative waitlist designs: Evidence from deceased donor kidneys," *Econometrica*, 89, 37–76.

Agarwal, N., A. Moehring, P. Rajpurkar, and T. Salz (2023): "Combining human expertise with artificial intelligence: Experimental evidence from radiology," Tech. rep., National Bureau of Economic Research.

Alvarez, R., A. J. Bonham, C. M. Buda, A. M. Carlin, A. A. Ghaferi, and O. A. Varban (2019): "Factors associated with long wait times for bariatric surgery," *Annals of surgery*, 270, 1103–1109.

Athey, S., L. K. Simon, O. N. Skans, J. Vikstrom, and Y. Yakymovych (2023): "The heterogeneous earnings impact of job loss across workers, establishments, and markets," *arXiv preprint arXiv:2307.06684*.

Athey, S., J. Tibshirani, and S. Wager (2019): "Generalized random forests," .

Battocchi, K., E. Dillon, M. Hei, G. Lewis, P. Oka, M. Oprescu, and V. Syrgkanis (2019): "EconML: A Python package for ML-Based heterogeneous treatment effects estimation," *Version 0. x*.

Bernstein, J., F. O. Roberts, B. B. Wiesel, and J. Ahn (2016): "Preoperative testing for hip fracture patients delays surgery, prolongs hospital stays, and rarely dictates care," *Journal of orthopaedic trauma*, 30, 78–80.

Brot-Goldberg, Z. C., A. Chandra, B. R. Handel, and J. T. Kolstad (2017): "What does a deductible do? The impact of cost-sharing on health care prices, quantities, and spending dynamics," *The Quarterly Journal of Economics*, 132, 1261–1318.

Cabral, M. and M. Dillender (2021): "Disparities in Health Care and Medical Evaluations by Gender: A Review of Evidence and Mechanisms," in *AEA Papers and Proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 111, 159–163.

Chan, D. C. and J. Gruber (2020): "Provider discretion and variation in resource allocation: The case of triage decisions," in *AEA papers and proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 110, 279–283.

Chandra, A., E. Flack, and Z. Obermeyer (2021): "The health costs of cost-sharing," Tech. rep., National Bureau of Economic Research.

Charlson, M. E., P. Pompei, K. L. Ales, and C. R. MacKenzie (1987): "A new method of classifying prognostic comorbidity in longitudinal studies: development and validation," *Journal of chronic diseases*, 40, 373–383.

Chen, C. L., S. D. McLeod, T. M. Lietman, H. Shen, W. J. Boscardin, H.-Y. P. Chang, M. A. Whooley, A. W. Gelb, S. J. Lee, and R. A. Dudley (2021): "Preoperative medical testing and falls in Medicare beneficiaries awaiting cataract surgery," *Ophthalmology*, 128, 208–215.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018a): "Double/debiased machine learning for treatment and structural parameters," .

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernandez-Val (2018b): "Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India," .

Coffman, J., T. Tran, T. Quast, M. S. Berlowitz, and S. H. Chae (2019): "Cost conscious care: preoperative evaluation by a cardiologist prior to low-risk procedures," *BMJ Open Quality*, 8, e000481.

Cullis, J. G., P. R. Jones, and C. Propper (2000): "Waiting lists and medical treatment: analysis and policies," *Handbook of health economics*, 1, 1201–1249.

DeCoster, C., K. C. Carriere, S. Peterson, R. Walld, and L. MacWilliam (1999): "Waiting times for surgical procedures," *Medical care*, 37, JS187–JS205.

Dossett, L. A., A. L. Edelman, G. Wilkinson, and S. M. Ruzycki (2022): "Reducing unnecessary preoperative testing," *bmj*, 379.

Doutreligne, M. and G. Varoquaux (2023): "How to Select Predictive Models for Decision Making or Causal Inference," *Available at SSRN 4467871*.

Dunn, M. (2018): "Who chooses part-time work and why," *Monthly Lab. Rev.*, 141, 1.

Einav, L., A. Finkelstein, and N. Mahoney (2022): "Producing health: Measuring value added of nursing homes," Tech. rep., National Bureau of Economic Research.

Einav, L., A. Finkelstein, S. Mullainathan, and Z. Obermeyer (2018): "Predictive modeling of US health care spending in late life," *Science*, 360, 1462–1465.

Garrido, M. V., A. Zentner, and R. Busse (2011): "The effects of gatekeeping: a systematic review of the literature," *Scandinavian journal of primary health care*, 29, 28–38.

Gasparini, A. (2018): "comorbidity: An R package for computing comorbidity scores," *Journal of Open Source Software*, 3, 648.

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

——— (2009): "Third degree waiting time discrimination: optimal allocation of a public sector healthcare treatment under rationing by waiting," *Health Economics*, 18, 977–986.

Gruber, J., B. R. Handel, S. H. Kina, and J. T. Kolstad (2020): "Managing intelligence: Skilled experts and AI in markets for complex products," Tech. rep., National Bureau of Economic Research.

Hacker, J. and D. Stanistreet (2004): "Equity in waiting times for two surgical specialties: a case study at a hospital in the North West of England," *Journal of Public Health*, 26, 56–60.

Hamilton, B. H., V. Ho, and D. P. Goldman (2000): "Queuing for surgery: is the US or Canada worse off?" *Review of Economics and Statistics*, 297–308.

Hartford, J., G. Lewis, K. Leyton-Brown, and M. Taddy (2017): "Deep IV: A flexible approach for counterfactual prediction," in *International Conference on Machine Learning*, PMLR, 1414–1423.

Hodge, W., T. Horsley, D. Albiani, J. Baryla, M. Belliveau, R. Buhrmann, M. O'Connor, J. Blair, and E. Lowcock (2007): "The consequences of waiting for cataract surgery: a systematic review," *Cmaj*, 176, 1285–1290.

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

Landi, S., E. Ivaldi, and A. Testi (2018): "Socioeconomic status and waiting times for health services: An international literature review and evidence from the Italian National Health System," *Health Policy*, 122, 334–351.

Lindsay, C. M. and B. Feigenbaum (1984): "Rationing by waiting lists," *The American economic review*, 74, 404–417.

Mahajan, D., I. Mitliagkas, B. Neal, and V. Syrgkanis (2022): "Empirical analysis of model selection for heterogenous causal effect estimation," *arXiv preprint arXiv:2211.01939*.

McDevitt, R. C. and J. W. Roberts (2014): "Market structure and gender disparity in health care: preferences, competition, and quality of care," *The RAND Journal of Economics*, 45, 116–139.

Moscelli, G., L. Siciliani, and V. Tonei (2016): "Do waiting times affect health outcomes? Evidence from coronary bypass," *Social science & medicine*, 161, 151–159.

Mullainathan, S. and Z. Obermeyer (2022): "Diagnosing physician error: A machine learning approach to low-value health care," *The Quarterly Journal of Economics*, 137, 679–727.

Nikolova, S., M. Harrison, and M. Sutton (2016): "The impact of waiting time on health gains from surgery: Evidence from a national patient-reported outcome dataset," *Health economics*, 25, 955–968.

Quan, H., V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J.-C. Luthi, L. D. Saunders, C. A. Beck, T. E. Feasby, and W. A. Ghali (2005): "Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data," *Medical care*, 1130–1139.

Ricci, W. M., G. J. Della Rocca, C. Combs, and J. Borrelli (2007): "The medical and economic impact of preoperative cardiac testing in elderly patients with hip fractures," *Injury*, 38, 49–52.

Shin, C. S., Y. N. Roh, and J. I. Kim (2014): "Delayed appendectomy versus early appendectomy in the treatment of acute appendicitis: a retrospective study," *World Journal of Emergency Surgery*, 9, 1–5.

Siciliani, L., M. Borowitz, and V. Moran (2013): *Waiting Time Policies in the Health Sector: What Works?*., OECD publishing.

Siciliani, L., V. Moran, and M. Borowitz (2014): "Measuring and comparing health care waiting times in OECD countries," *Health policy*, 118, 292–303.

Smirthwaite, G., M. Lundström, S. Albrecht, and K. Swahnberg (2014): "Indication criteria for cataract extraction and gender differences in waiting time," *Acta ophthalmologica*, 92, 432–438.

Sobolev, B. and G. Fradet (2008): "Delays for coronary artery bypass surgery: how long is too long?" *Expert review of pharmacoeconomics & outcomes research*, 8, 27–32.

Sobolev, B. G., A. R. Levy, L. Kuramoto, R. Hayden, and J. M. FitzGerald (2006): "Do longer delays for coronary artery bypass surgery contribute to preoperative mortality in less urgent patients?" *Medical care*, 680–686.

Sripa, P., B. Hayhoe, P. Garg, A. Majeed, and G. Greenfield (2019): "Impact of GP gatekeeping on quality of care, and health outcomes, use, and expenditure: a systematic review," *British Journal of General Practice*, 69, e294–e303.

Syrgkanis, V., V. Lei, M. Oprescu, M. Hei, K. Battocchi, and G. Lewis (2019): "Machine learning estimation of heterogeneous treatment effects with instruments," *Advances in Neural Information Processing Systems*, 32.

Vaidya, V., G. Partha, and M. Karmakar (2012): "Gender differences in utilization of preventive care services in the United States," *Journal of women's health*, 21, 140–145.

Viberg, N., B. C. Forsberg, M. Borowitz, and R. Molin (2013): "International comparisons of waiting times in health care–Limitations and prospects," *Health policy*, 112, 53–61.

Wei, J., L. Zeng, S. Li, F. Luo, Z. Xiang, and Q. Ding (2019): "Relationship between comorbidities and treatment decision-making in elderly hip fracture patients," *Aging clinical and experimental research*, 31, 1735–1741.

Part I

Appendix

Table of Contents

A Data, Surgery Description and Sample Selection

A.1 Data

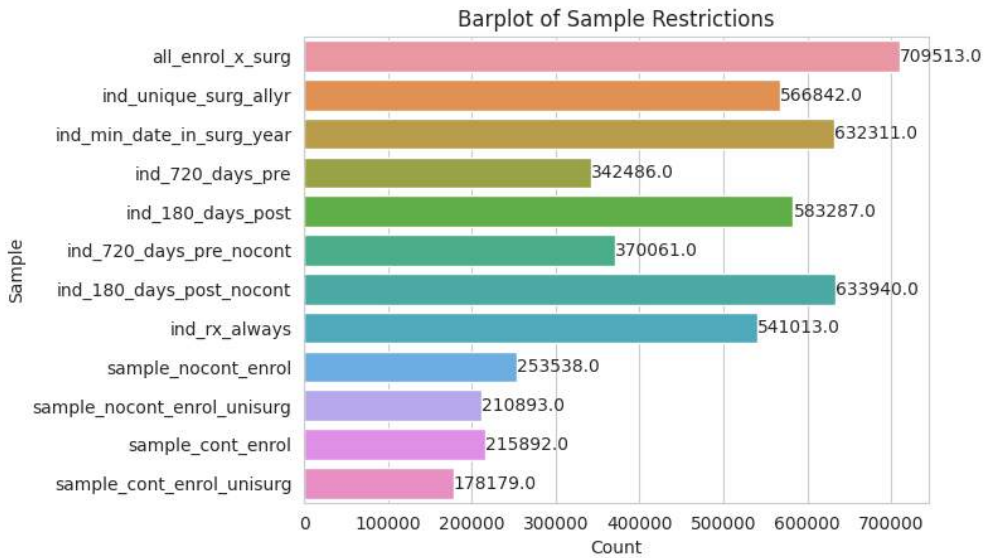


Figure A1. Restrictiveness of Sample Restrictions

Notes:

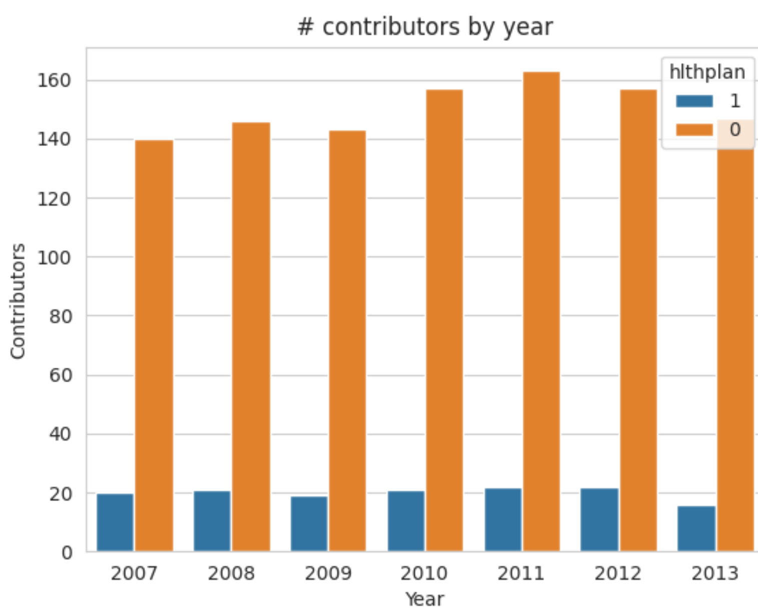


Figure A2. Data Sources: Employers and Health Plans

Notes:

A.2 Surgeries Description

Table A1. Surgery Descriptions, CPT 4 and ICD 9 Codes

Surgery	Description	CPT4 Code	ICD9 Code
Appendectomy	Appendectomy is a surgical procedure to remove the appendix, a small, finger-like pouch that hangs off the large intestine, usually due to inflammation or infection.	44950, 44955, 44960, 44970	4701, 4709, 4711, 4719
Cataract	Cataract surgery is a procedure to remove the lens of the eye that has become cloudy, usually due to age, and replace it with an artificial lens.	66830, 66840, 66850, 66852, 66920, 66930, 66940, 66982, 66983, 66984	1371
Cholecystectomy	Cholecystectomy is a surgical procedure to remove the gallbladder, an organ located near the liver that stores bile, usually due to inflammation, infection, or the presence of gallstones.	47480, 47562, 47563, 47564, 47570, 47600, 47605, 47610, 47612, 47620, 47720, 47721, 47740, 47741	5123
Colectomy	Colectomy is a surgical procedure to remove part or all of the colon, usually due to cancer, inflammatory bowel disease, or other conditions.	44140, 44141, 44143, 44144, 44145, 44146, 44147, 44150, 44151, 44155, 44156, 44157, 44158, 44160, 44204, 44205, 44206, 44207, 44208, 44210, 44211, 44212, 44213	4573, 4575, 1733, 1735
Coronary Bypass	Coronary bypass surgery is a procedure to create a new route for blood to flow around blocked or narrowed arteries in the heart, usually due to coronary artery disease.	33510, 33511, 33512, 33513, 33514, 33516, 33533, 33534, 33535, 33536	3611, 3612, 3613, 3614
Endarterectomy	Endarterectomy is a surgical procedure to remove plaque or fatty deposits from the lining of an artery, usually in the neck or legs, to improve blood flow.	35301, 35390	3810, 3811, 3812, 3813, 3814, 3815, 3816, 3817, 3818

Table A2. Surgery Descriptions, CPT 4 and ICD 9 Codes

Surgery	Description	CPT4 Code	ICD9 Code
Hip Replacement	Hip replacement surgery is a procedure to remove a damaged or diseased hip joint and replace it with an artificial joint made of metal, plastic, or ceramic.	27125, 27130	8151
Hysterectomy	Hysterectomy is a surgical procedure to remove the uterus, usually due to cancer, fibroids, or other conditions.	58150, 58152, 58180, 58200, 58210, 58240, 58241, 58242, 58243, 58244, 58548, 58550, 58552, 58553, 58554, 58570, 58571, 58572, 58573	6849, 6859, 6851, 6841, 6839, 6831, 6869, 6861
Inguinal Hernia	Inguinal hernia repair is a surgical procedure to repair a hernia in the lower abdomen or groin area, usually due to a weakness in the abdominal wall.	49491, 49492, 49495, 49496, 49500, 49501, 49505, 49507, 49520, 49521, 49525	5300, 5301, 5302, 5303, 5304, 5305, 5310, 5311, 5312, 5313, 5314, 5315, 5316, 5317
Kidney Transplant	Kidney transplant surgery is a procedure to replace a diseased or damaged kidney with a healthy kidney from a donor.	50360, 50365, 50380	5569
Knee Replacement	Knee replacement surgery replaces a damaged knee joint with an artificial joint to relieve pain and improve mobility, usually due to arthritis or injury.	27438, 27440, 27441, 27442, 27443, 27444, 27445, 27446, 27447	8154
Liver Transplant	Liver transplant surgery is a procedure to replace a diseased or damaged liver with a healthy liver from a donor.	47135	5059
Mastectomy	Mastectomy is a surgical procedure to remove one or both breasts, usually due to breast cancer.	19300, 19301, 19302, 19303, 19305, 19306, 19307	8541, 8542, 8543, 8544
Nephrostomy	Nephrostomy is a surgical procedure to create a temporary or permanent opening between the kidney and the skin to allow urine to drain out of the body, usually due to a blockage or obstruction in the urinary system.	50040	5501, 5502, 5503, 5504
Prostatectomy	Prostatectomy is a surgical procedure to remove all or part of the prostate gland, usually due to prostate cancer or an enlarged prostate.	55810, 55812, 55815, 55821, 55831, 55840, 55842, 55845, 55866	603, 604, 605, 6029, 6069, 6021, 6062
Splenectomy	Splenectomy is a surgical procedure to remove the spleen, usually due to a ruptured spleen, cancer, or other conditions.	38100, 38101, 38102, 38115, 38120	415
Thyroidectomy	Thyroidectomy is a surgical procedure to remove all or part of the thyroid gland, usually due to thyroid cancer, hyperthyroidism, or other conditions.	60240, 60252, 60254, 60260, 60270, 60271	64, 639

A.3 Plan Types Description

Table A3. Marketscan Type of Plans and Characteristics

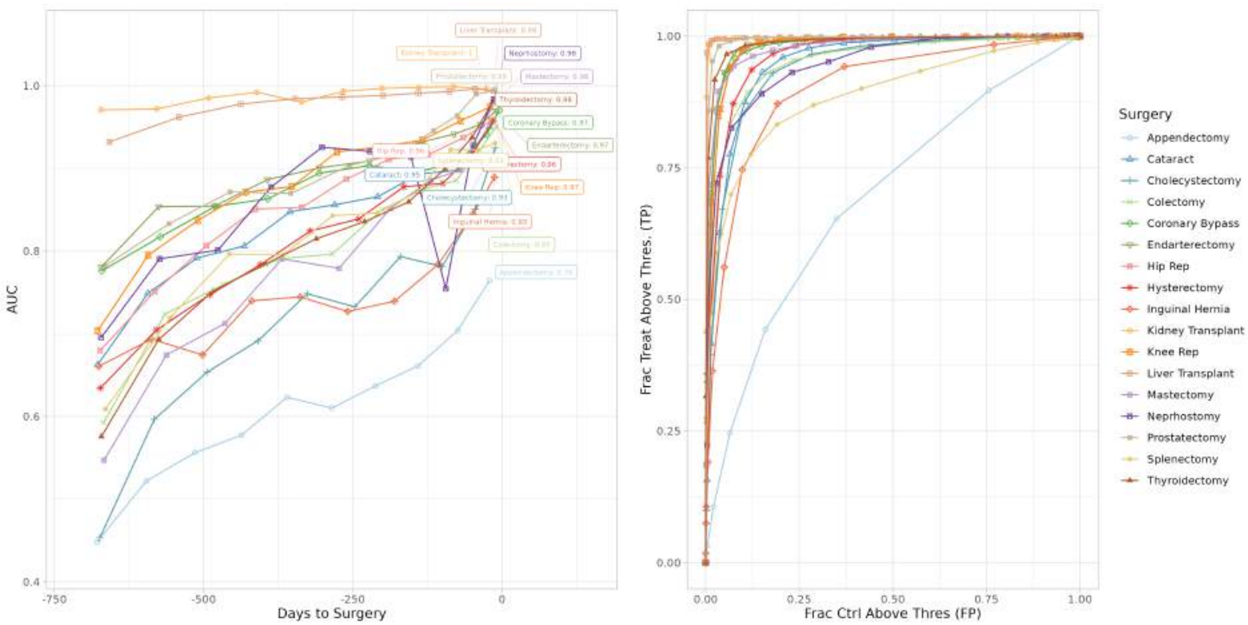
Plan Type	Patient incentive to use certain providers?	PCP assigned?	Referrals from PCP to specialist required?	Out of network services covered?	Partially or Fully capitated?
Basic/ Major Medical	No	No	n/a	n/a	No
Comprehensive	No	No	n/a	n/a	No
Exclusive Provider Organization (**EPO**)	Yes	Yes	Yes	No	No
Health Management Organization (**HMO**)	Yes	Yes	Yes	No	Yes
Non-Capitated Point-of-Service (**POS**)	Yes	Yes	Yes	Yes	No
Preferred Provider Organization (**PPO**)	Yes	No	n/a	Yes	No
Capitated or Partially Capitated Point-of-Service (**POSCap**)	Yes	Yes	Yes	Yes	Yes
Consumer-Driven Health Plan (**CDHP**)	Varies	No	n/a	Varies	No
High Deductible Health Plan (**HDHP**)	Varies	No	n/a	Varies	No

B Classifiers Training, Performance and Model Selection

B.1 Training

Preprocessing. Describe data formatting for prediction task

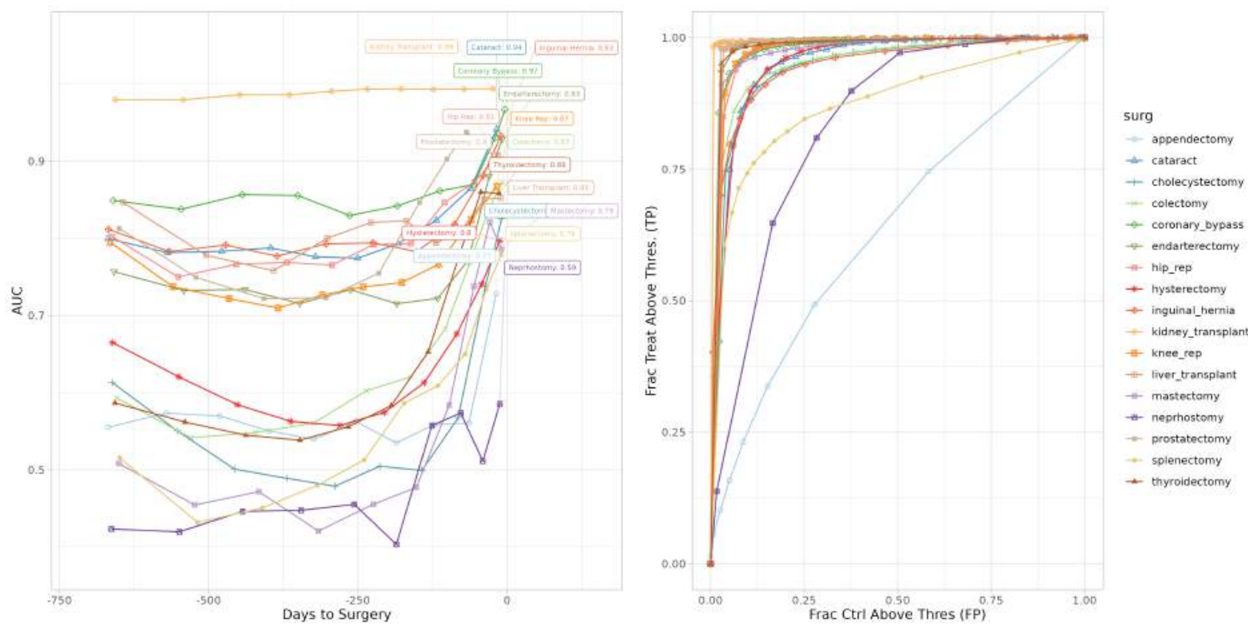
B.2 Model Selection



(a) AUC Over Time (b) AUC Overall

Figure B3. LASSO-Logistic Regression Performance, Over Time and Overall

Notes: Panel (a) shows the average AUC scores at various points in patients' medical trajectories before surgery. The scores are computed on a hold-out sample of treated and control patients. Treated patients visits are split in deciles of how far back in time they occurred before the surgery, and control patient visits are just randomly selected. AUC are computed over each of these deciles. Panel (b) displays the TP/FP graph of the entire method, computed for every possible threshold between 1 (bottom left corner) and 0 (top right) with steps of 0.05.

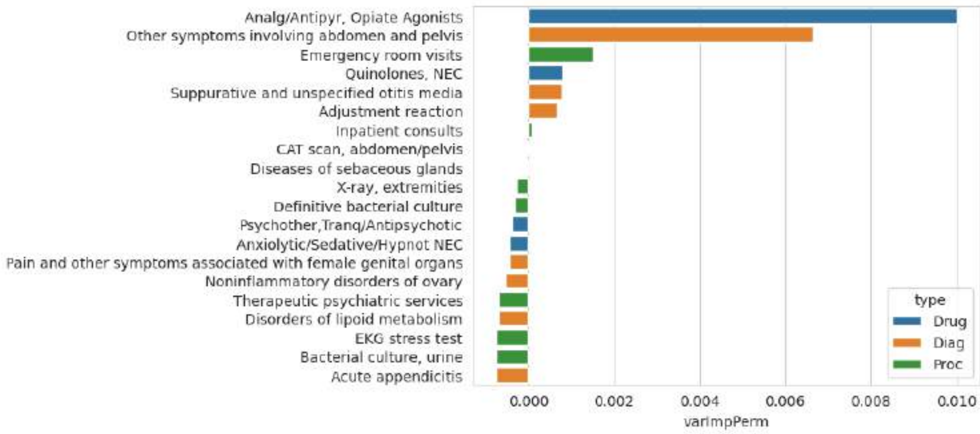


(a) AUC Over Time (b) AUC Overall

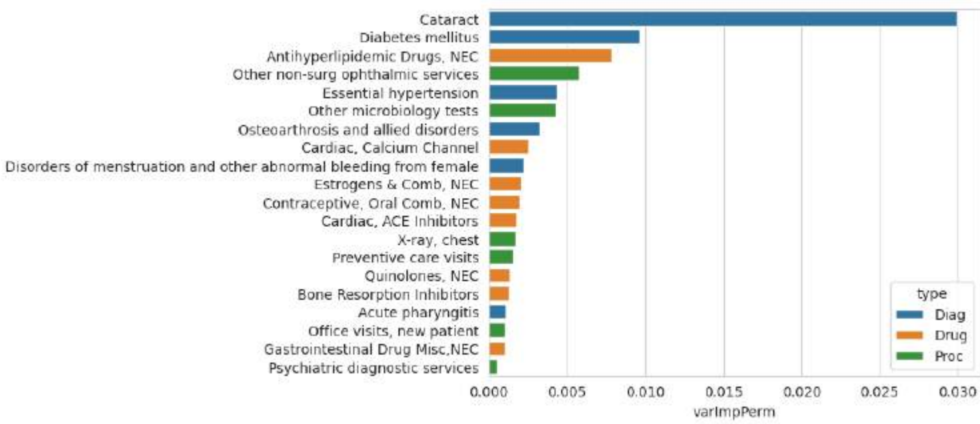
Figure B4. LSTM Regression Performance, Over Time and Overall

Notes: Panel (a) shows the average AUC scores at various points in patients' medical trajectories before surgery. The scores are computed on a hold-out sample of treated and control patients. Treated patients visits are split in deciles of how far back in time they occurred before the surgery, and control patient visits are just randomly selected. AUC are computed over each of these deciles. Panel (b) displays the TP/FP graph of the entire method, computed for every possible threshold between 1 (bottom left corner) and 0 (top right) with steps of 0.05.

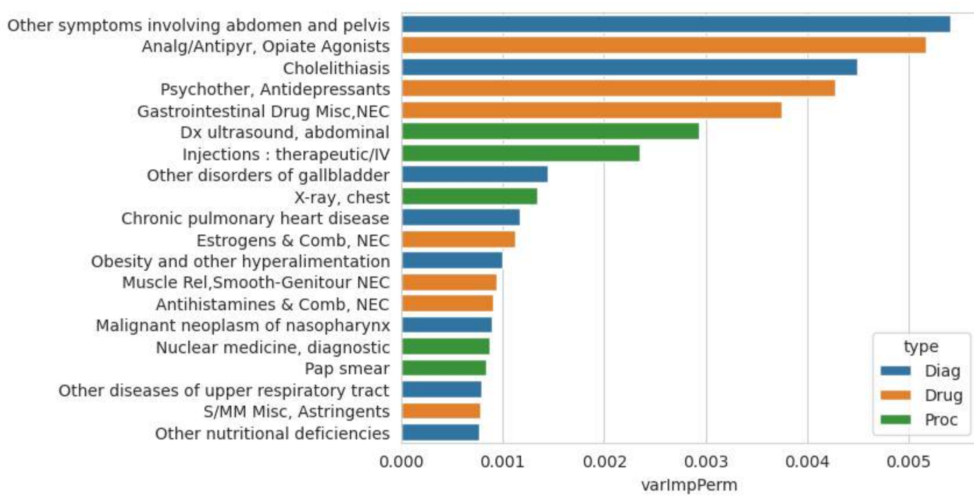
B.3 Interpretation: Variable Importance Plots



(a) Appendectomy



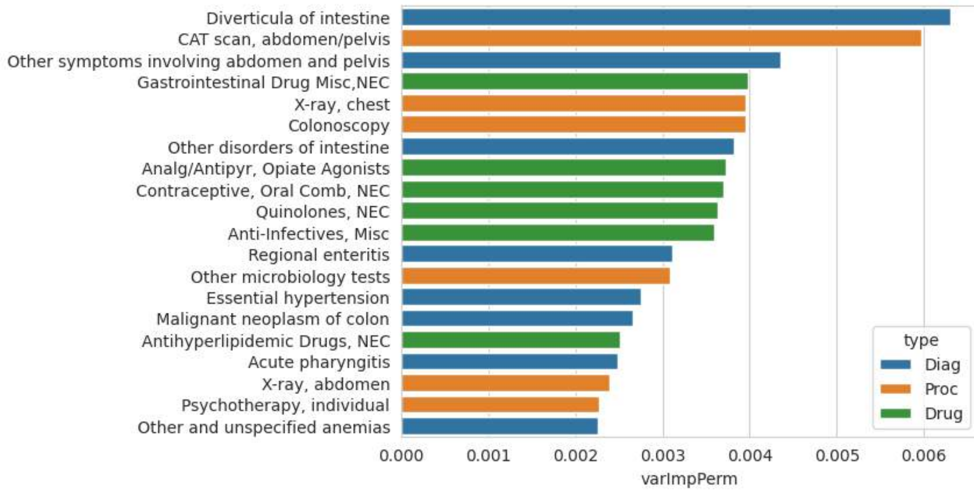
(b) Cataract



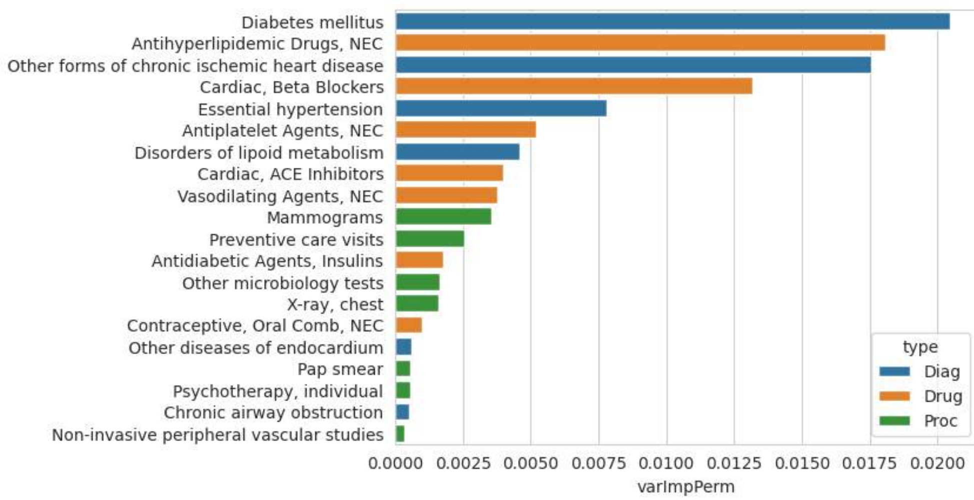
(c) Cholecystectomy

Figure B5. Permutation Variable Importance

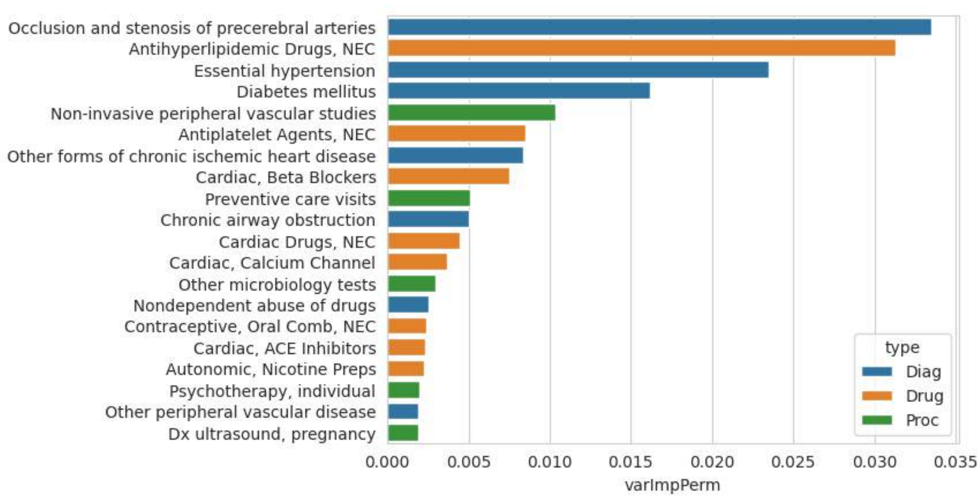
Notes: Each panel shows the 20 variables among Diagnoses, Drugs and Procedures with the highest permutation importance in the test set.



(a) Colectomy



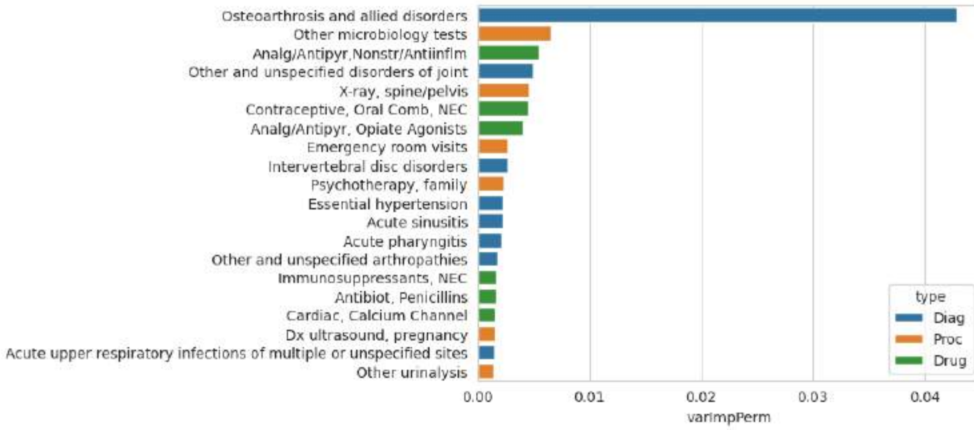
(b) Coronary Bypass



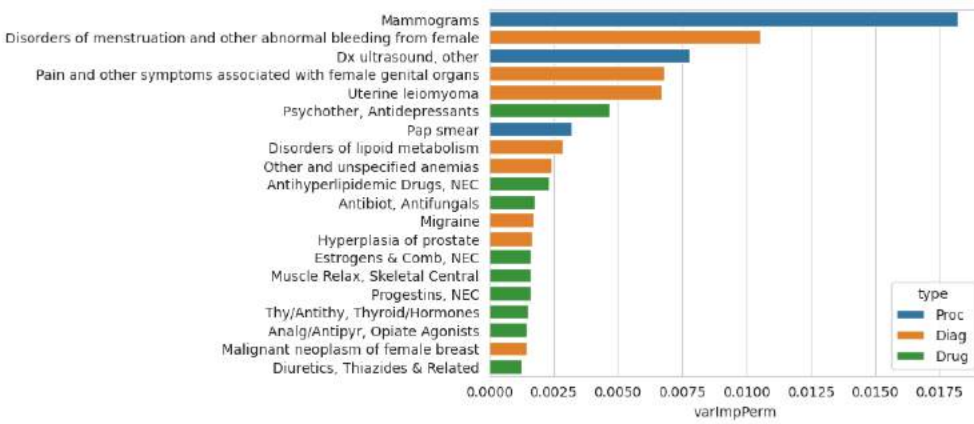
(c) Endarterectomy

Figure B6. Permutation Variable Importance

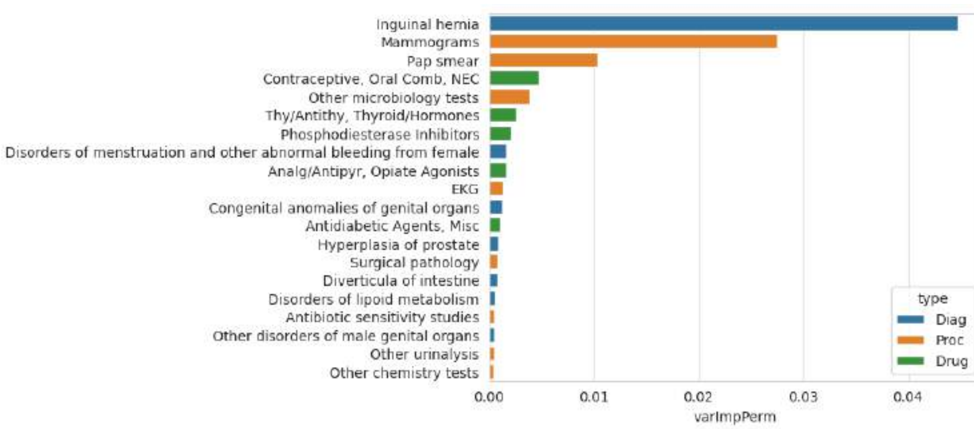
Notes: Each panel shows the 20 variables among Diagnoses, Drugs and Procedures with the highest permutation importance in the test set.



(a) Hip Replacement



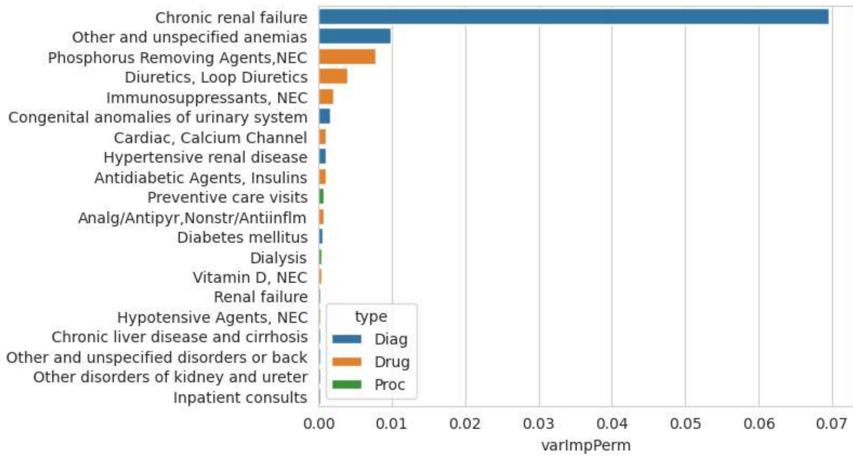
(b) Hysterectomy



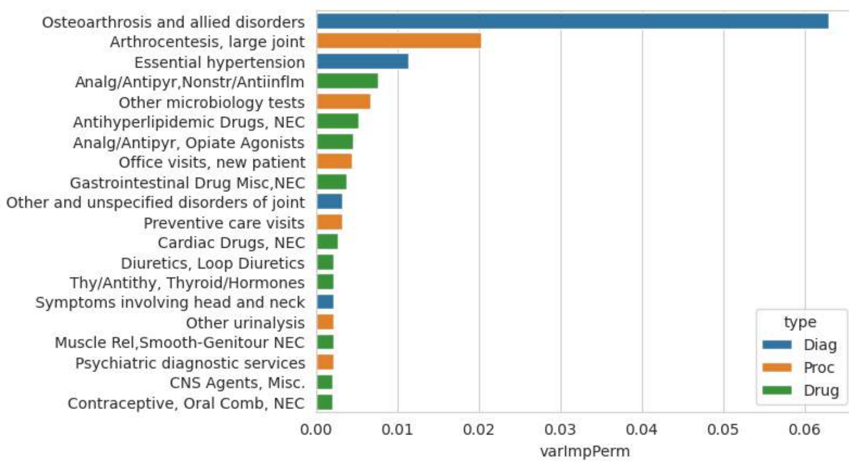
(c) Inguinal Hernia

Figure B7. Permutation Variable Importance

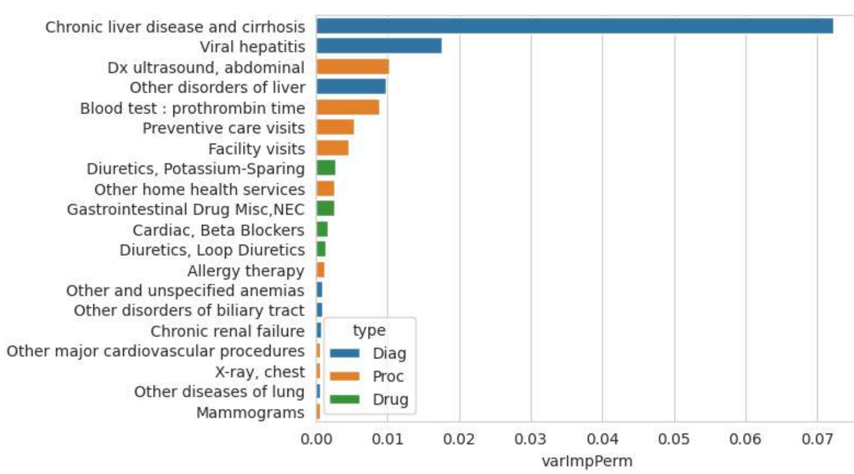
Notes: Each panel shows the 20 variables among Diagnoses, Drugs and Procedures with the highest permutation importance in the test set.



(a) Kidney Trasnplant



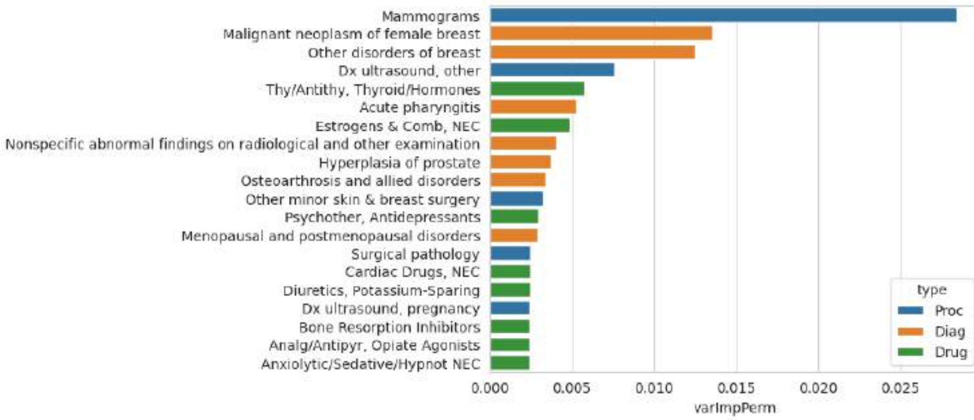
(b) Knee Replacement



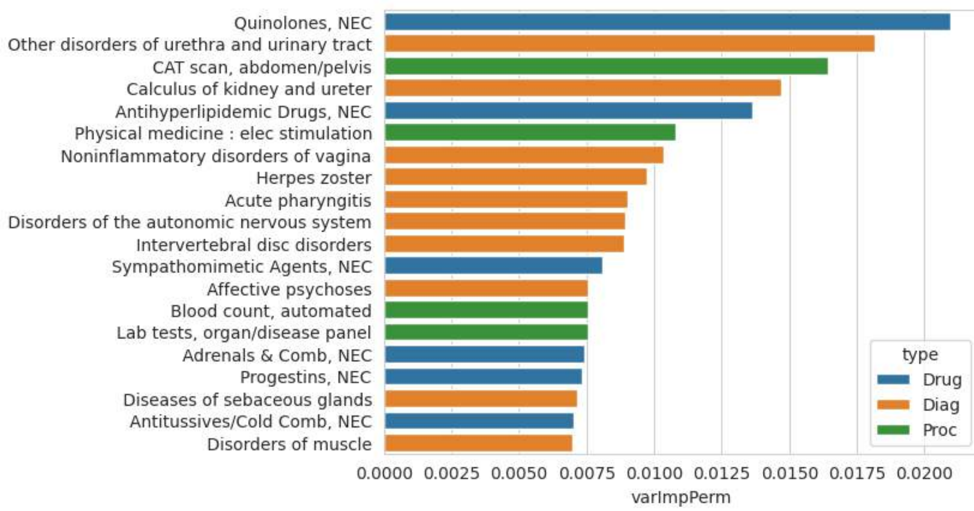
(c) Liver Transplant

Figure B8. Permutation Variable Importance

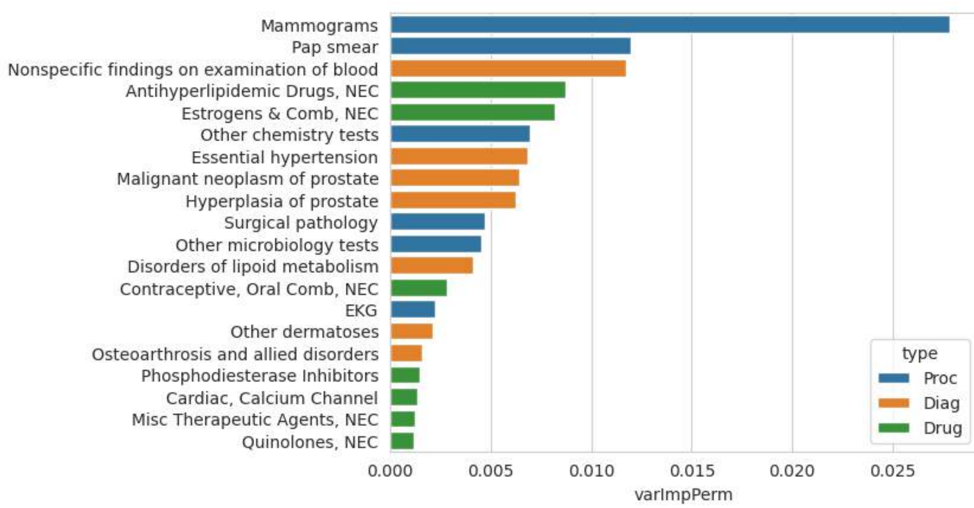
Notes: Each panel shows the 20 variables among Diagnoses, Drugs and Procedures with the highest permutation importance in the test set.



(a) Mastectomy



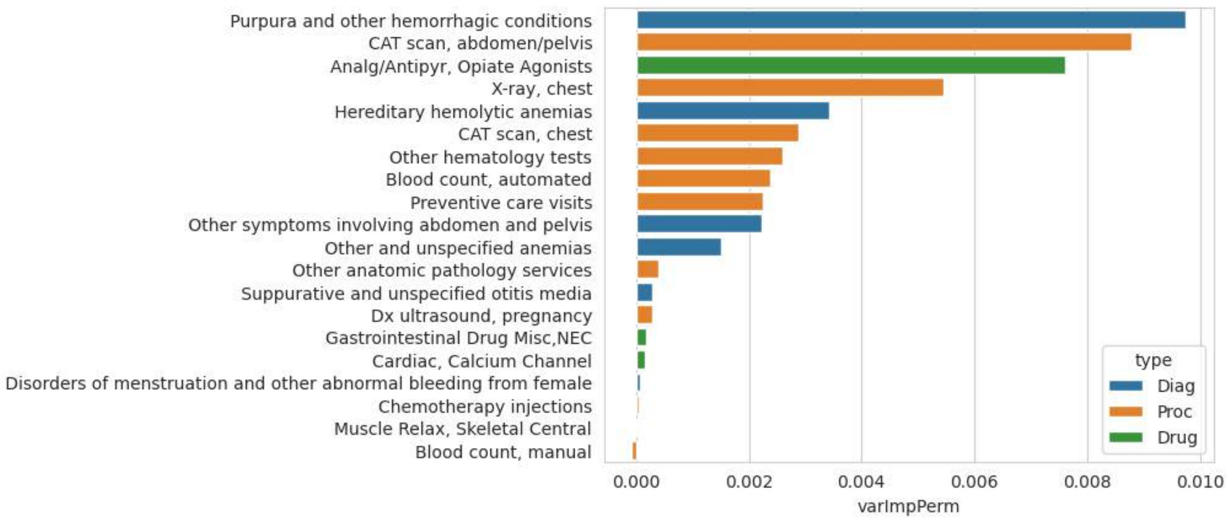
(b) Nephrostomy



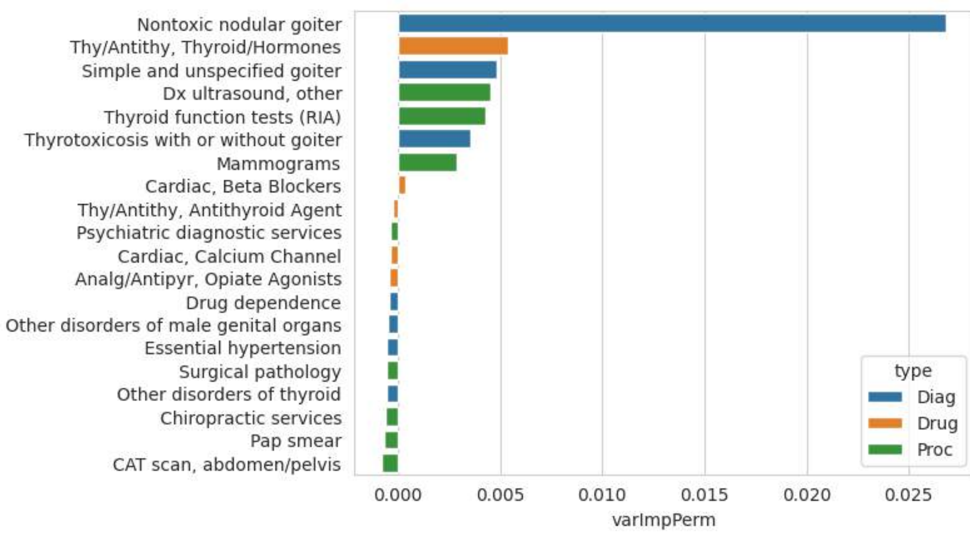
(c) Prostatectomy

Figure B9. Permutation Variable Importance

Notes: Each panel shows the 20 variables among Diagnoses, Drugs and Procedures with the highest permutation importance in the test set.



(a) Splenectomy



(b) Thyroidectomy

Figure B10. Permutation Variable Importance

Notes: Each panel shows the 20 variables among Diagnoses, Drugs and Procedures with the highest permutation importance in the test set.

C Additional Analyses

C.1 Detection to Treatment

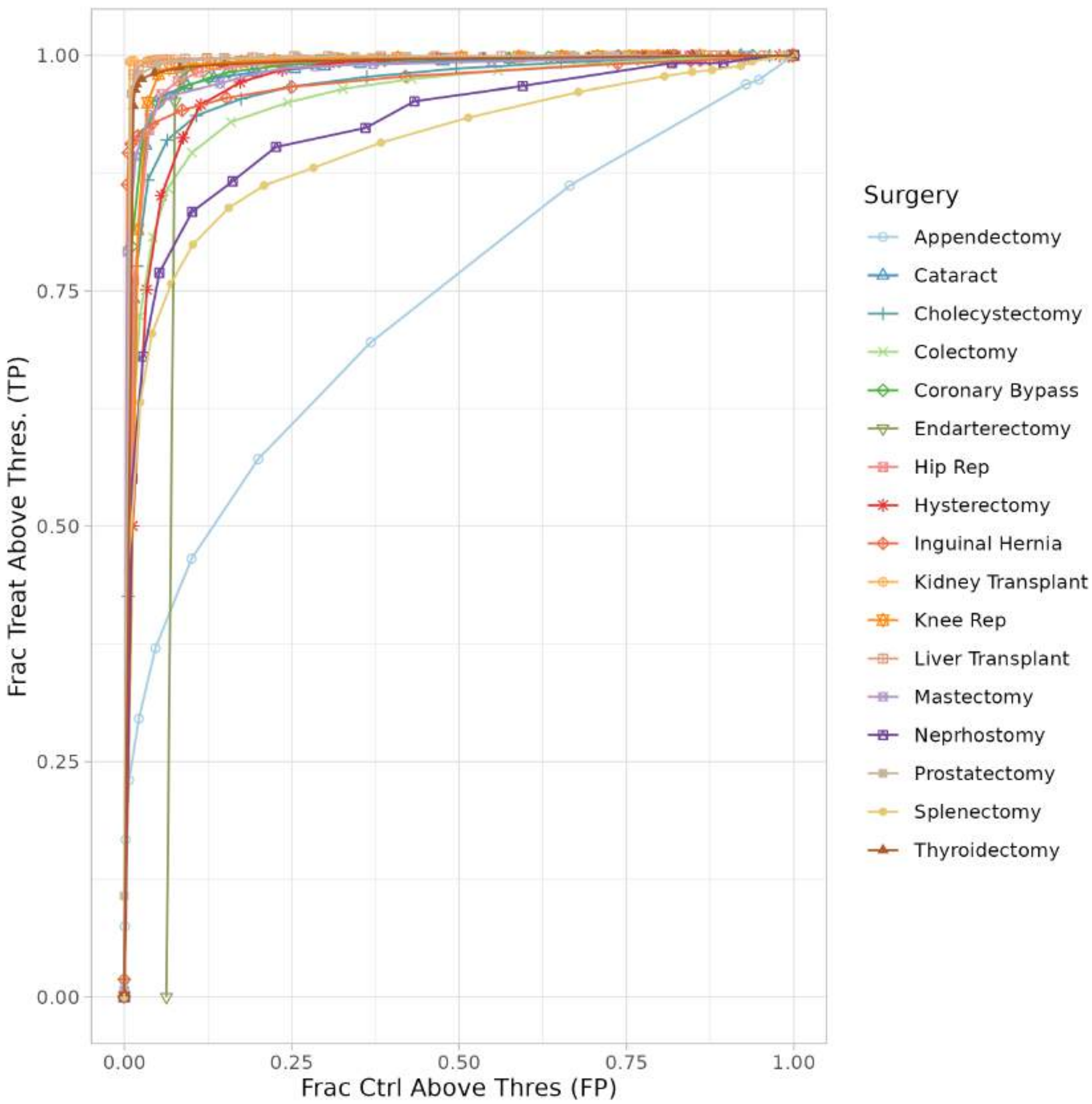


Figure C11. AUC with Full Medical History

Notes: The figure displays the TP/FP graph of the entire method, computed for every possible threshold between 1 (bottom left corner) and 0 (top right) with steps of 0.05.

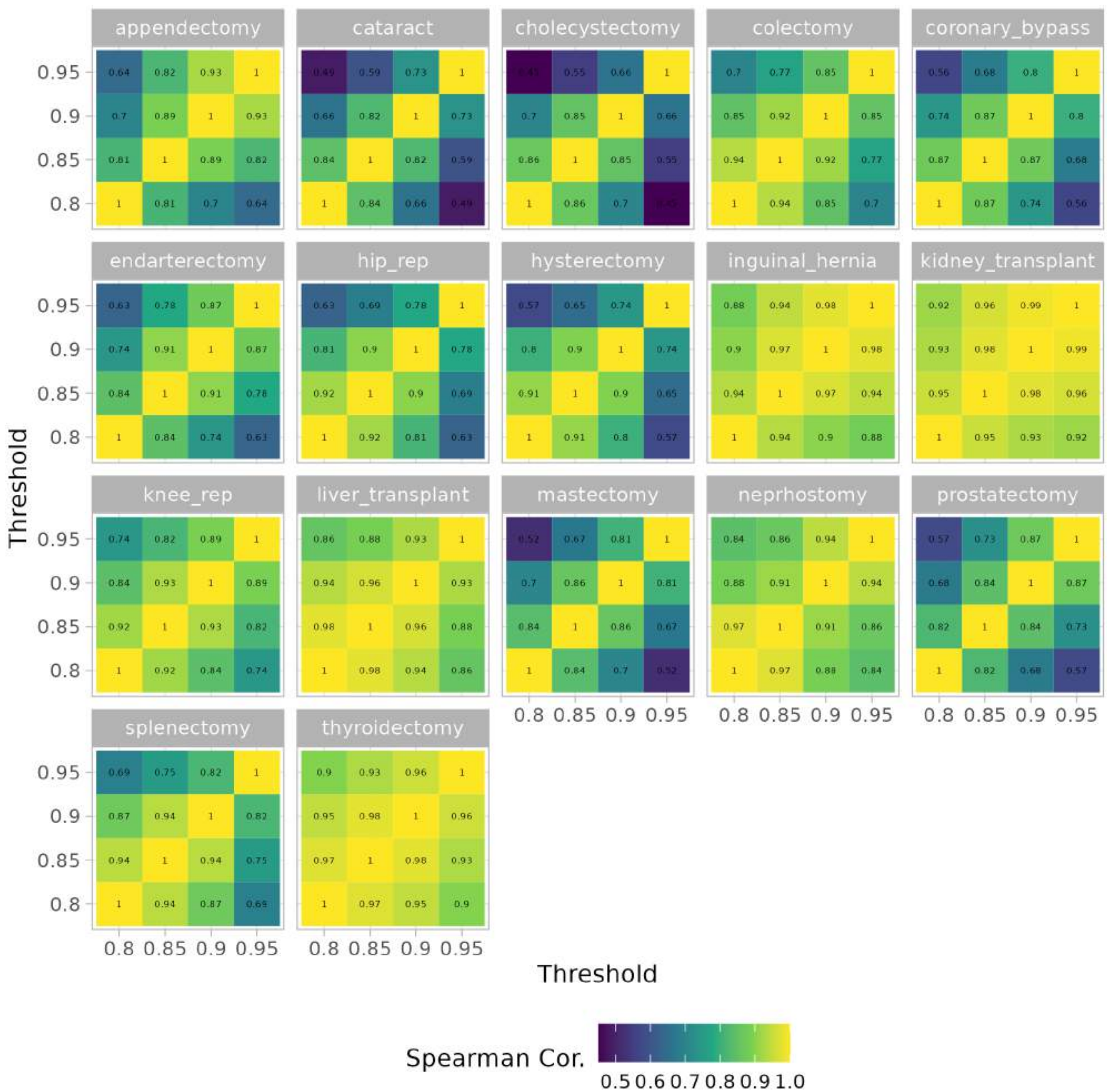
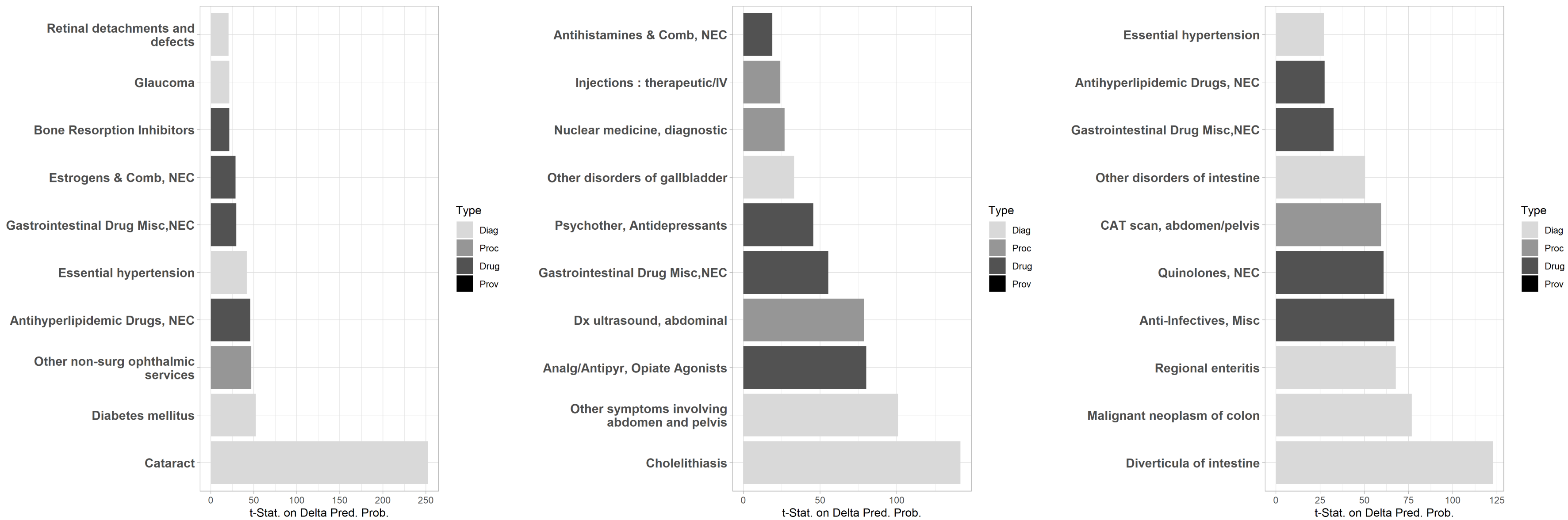


Figure C12. Wait Times Correlations Across Detection Thresholds

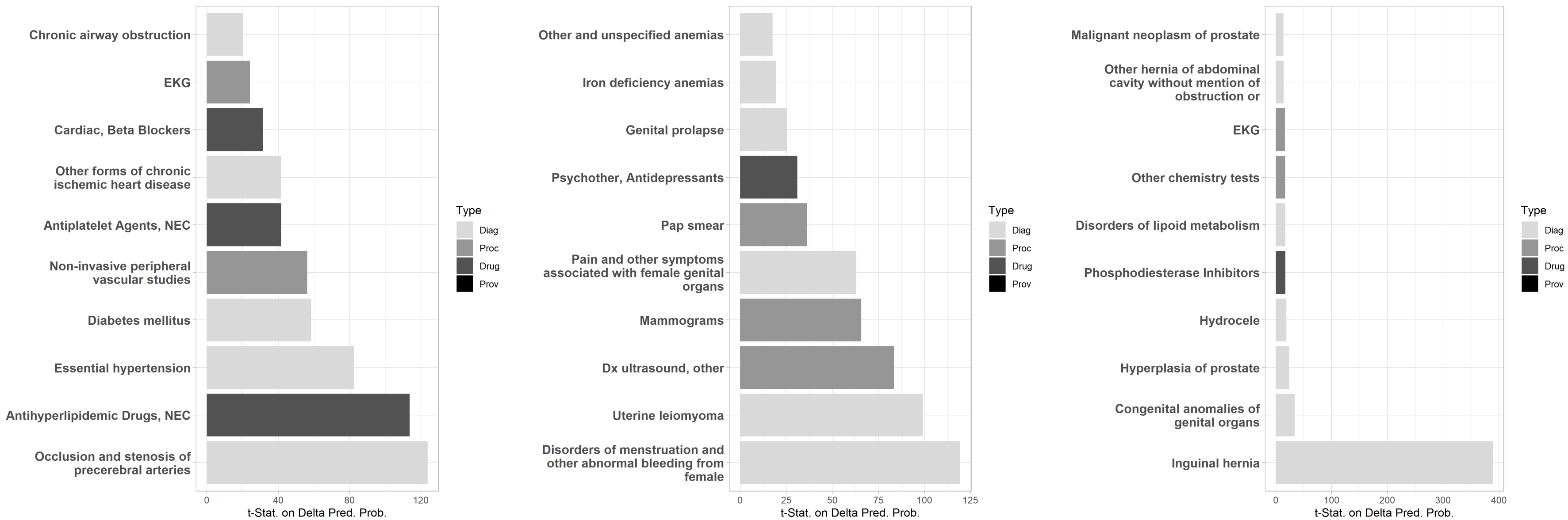
Notes: Each panel shows the Spearman (rank) correlation between wait times implied by detection thresholds ranging from 0.8 to 0.95 for all the surgeries in the sample.



(a) Cataract (b) Cholecytectomy (c) Colectomy

Figure C13. Delta Probability Variable Importance

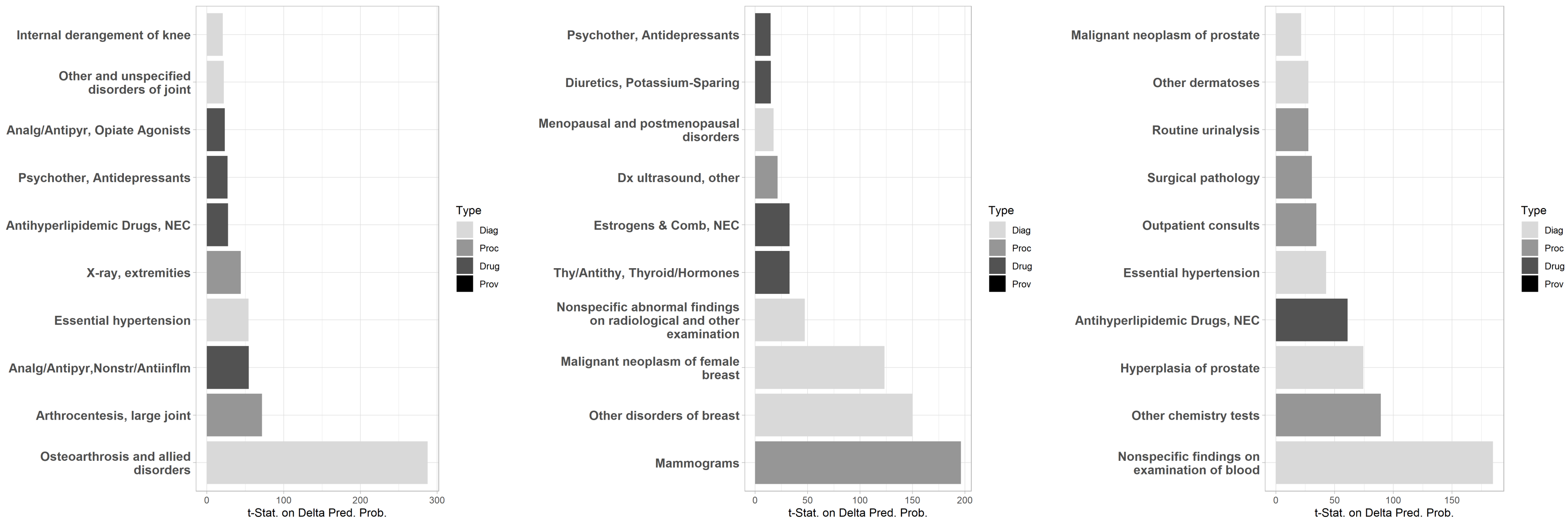
Notes: We illustrate the key medical "events" that underlie our prediction of surgical need. In each panel, we list the top 10 events—including procedures, drugs taken, and diagnoses— that suggest the patient will receive the focal surgery in the future. The events are selected and ordered in the figure using the t-statistic we find from a regression of (a) the change in predicted probability of surgery between two visits, $\Delta\hat{p}_{ist}$, on (b) the presence of the event between these visits.



(a) Endarterectomy (b) Hysterectomy (c) Inguinal Hernia

Figure C14. Delta Probability Variable Importance

Notes: We illustrate the key medical "events" that underlie our prediction of surgical need. In each panel, we list the top 10 events—including procedures, drugs taken, and diagnoses— that suggest the patient will receive the focal surgery in the future. The events are selected and ordered in the figure using the t-statistic we find from a regression of (a) the change in predicted probability of surgery between two visits, $\Delta\hat{p}_{ist}$, on (b) the presence of the event between these visits.



(a) Knee Replacement

(b) Mastectomy

(c) Prostatectomy

Figure C15. Delta Probability Variable Importance

Notes: We illustrate the key medical "events" that underlie our prediction of surgical need. In each panel, we list the top 10 events—including procedures, drugs taken, and diagnoses— that suggest the patient will receive the focal surgery in the future. The events are selected and ordered in the figure using the t-statistic we find from a regression of (a) the change in predicted probability of surgery between two visits, $\Delta\hat{p}_{ist}$, on (b) the presence of the event between these visits.

Table C4. Performance Relative to Deterministic Method

	True Positive Rate			1 - False Positive Rate		
Surgery	Detection	Deterministic	Difference	Detection	Deterministic	Difference
Cataract	0.91	0.95	-0.03	0.97	0.98	-0.01
Cholecystectomy	0.76	0.68	0.09	0.98	0.99	-0.01
Colectomy	0.71	0.41	0.30	0.98	0.98	-0.00
Coronary Bypass	0.91	0.85	0.07	0.97	0.87	0.10
Endarterectomy	0.98	0.97	0.00	0.98	0.99	-0.02
Hip Rep	0.92	0.94	-0.01	0.96	0.96	0.01
Hysterectomy	0.76	0.63	0.13	0.97	0.94	0.03
Inguinal Hernia	0.90	0.90	-0.01	0.99	0.99	-0.00
Knee Rep	0.94	0.98	-0.03	0.96	0.95	0.01
Mastectomy	0.90	0.96	-0.06	0.98	0.81	0.17
Prostatectomy	0.98	0.91	0.07	0.97	0.95	0.03
Splenectomy	0.63	0.22	0.41	0.98	0.99	-0.02

Notes: This table compares (1) our wait time measure, labeled *Detection*, computed as the difference between (a) the first visit when models predict the patient will require the surgery with above 90% probability and (b) the day and the surgery and (2) a measure labeled *Deterministic* computed as the difference between (a) the first visit when the most predictive medical event occurs (b) the day and the surgery. For each measure and each surgery we compute the fraction of surgical patients assigned a wait time (true positive rate) and the fraction of non-surgical patients incorrectly assigned a wait time (false positive rate). Most predictive events are identified as the regressor with the highest t-Stat in regression 3 estimated surgery by surgery.

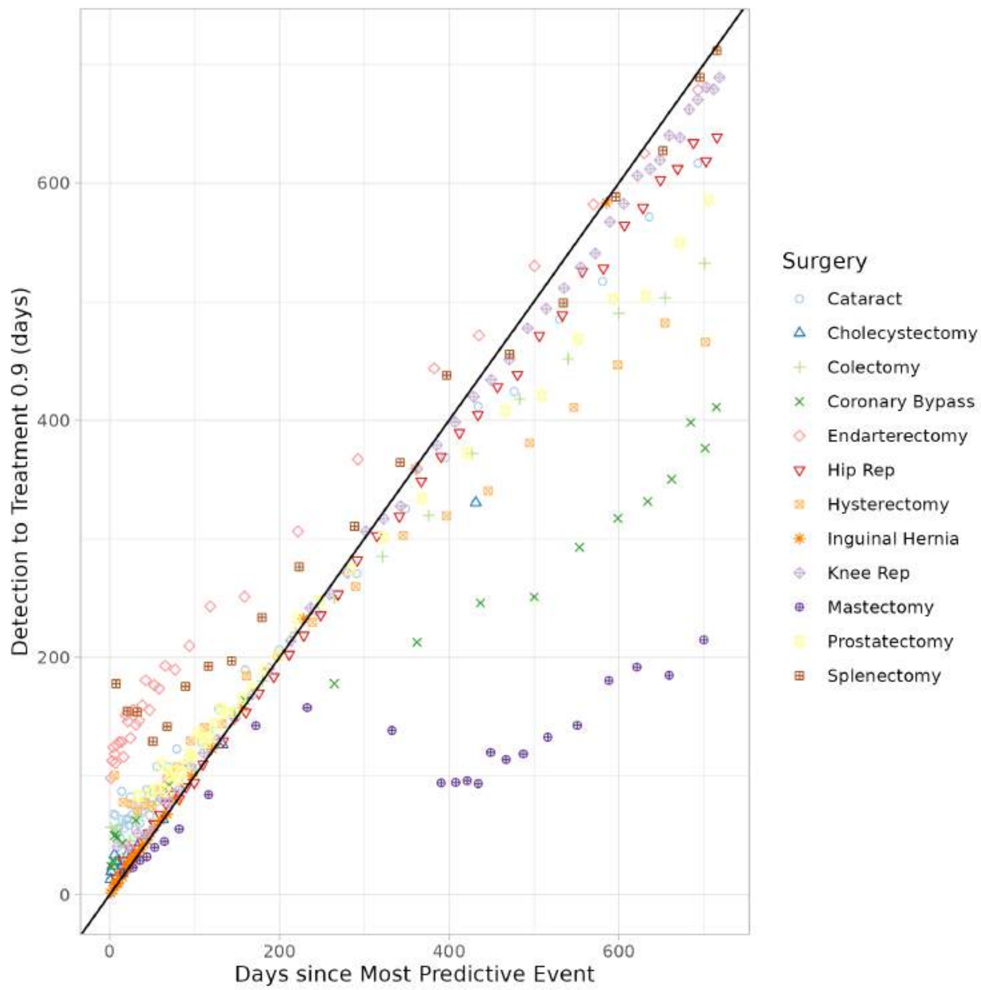
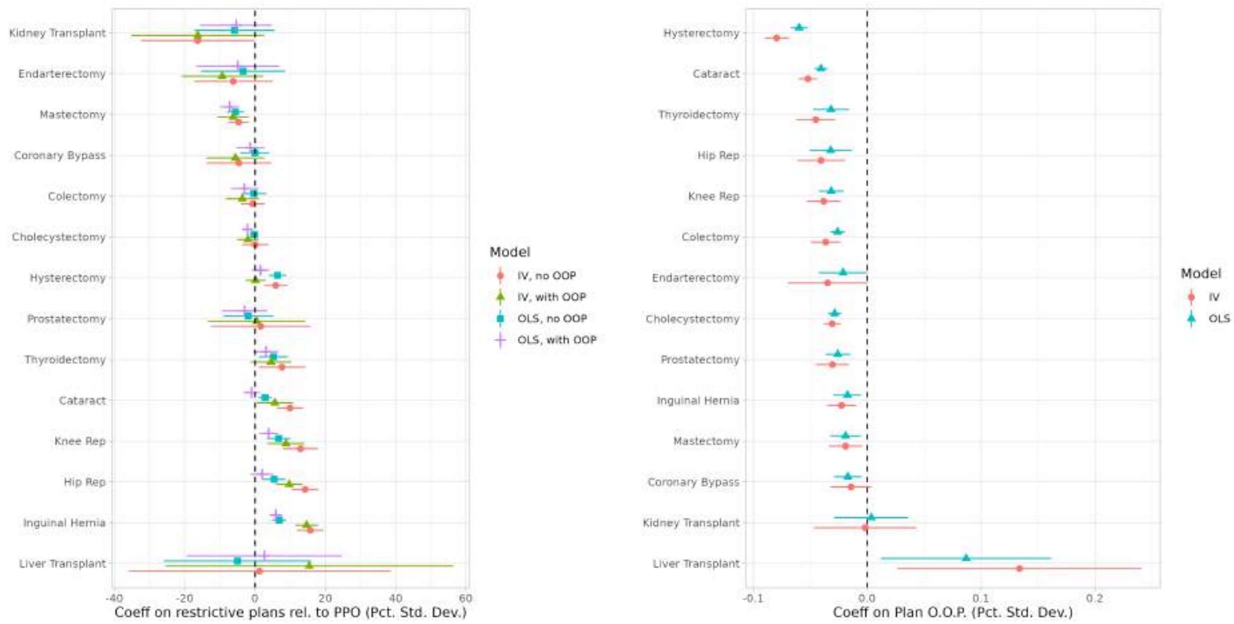


Figure C16. Detection to Treatment Compared to Deterministic

Notes: This Figure compares (1) our wait time measure, *Detection*, computed as the difference between (a) the first visit when models predict the patient will require the surgery with above 90% probability and (b) the day and the surgery and (2) a measure computed as the difference between (a) the first visit when the most predictive medical event occurs (b) the day and the surgery. Most predictive events are identified as the regressor with the highest t-Stat in regression 3 estimated surgery by surgery.

C.2 Wait Times and Health Insurance Design



(a) HMO Delays by Surgery

(b) Out-of-Pocket Impact on Wait Times

Figure C17. Insurance Design and Wait Times

Notes: Left Panel shows the estimated OLS and IV coefficients and 95% confidence intervals for HMOs relative to PPOs. The regressions alternatively include or exclude control (instrumented) for plan OOP share, to show both the total effect of HMO plans and the effect of plan design net of cost-sharing. Right panel shows estimates and confidence intervals on plan average OOP share. In both panels, regressions are run surgery by surgery and include patient controls, month and year of surgery fixed effects. Sample includes patients from employers only and excludes healthplan contributors. Standard errors are robust .

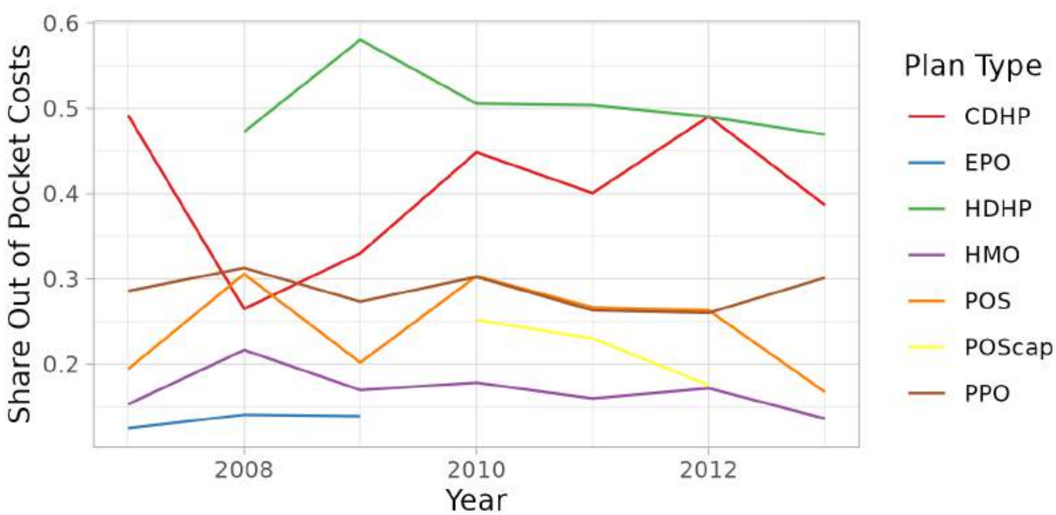
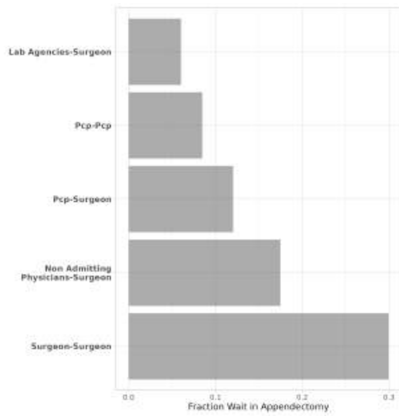
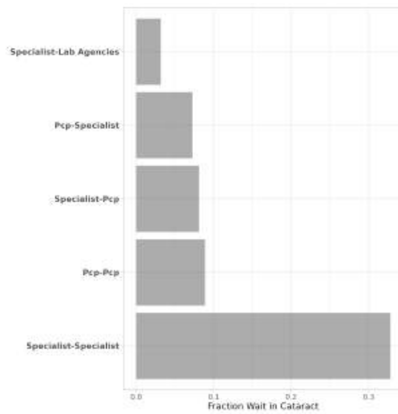


Figure C18. Out-of-Pocket Share by Insurance Plan Type

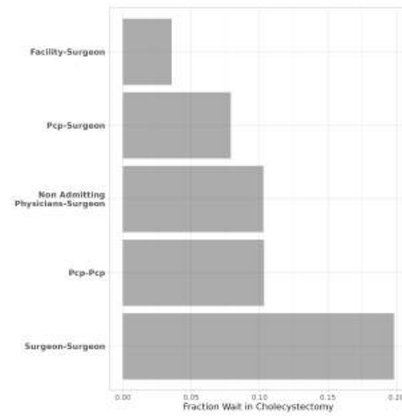
Notes: The Figure shows the average out-of-pocket share by type of plan, where a plan is defined as a contributor by plan type.



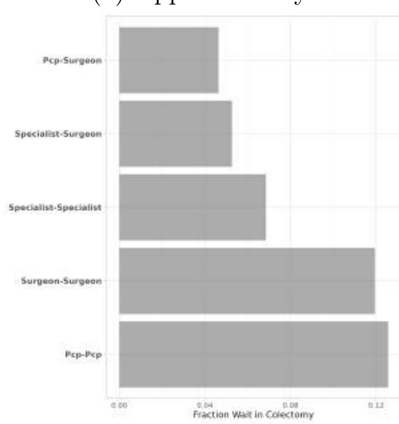
(a) Appendectomy



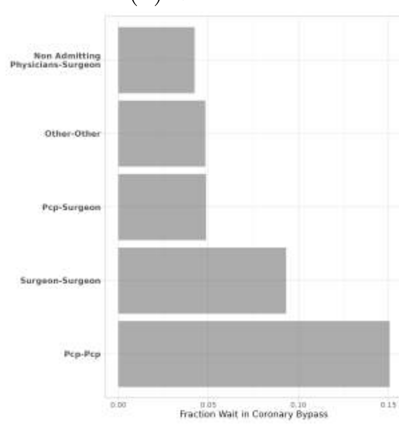
(b) Cataract



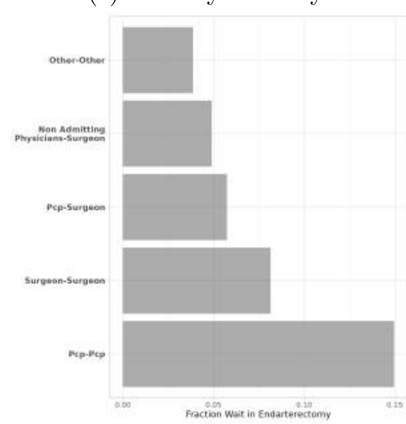
(c) Cholecystectomy



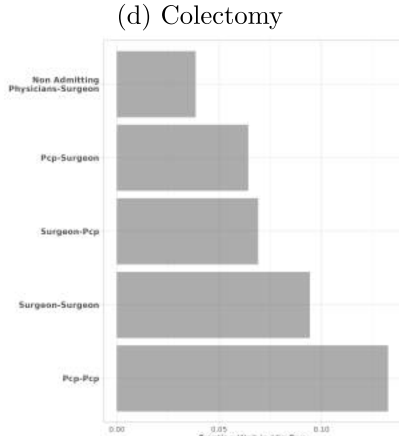
(d) Colectomy



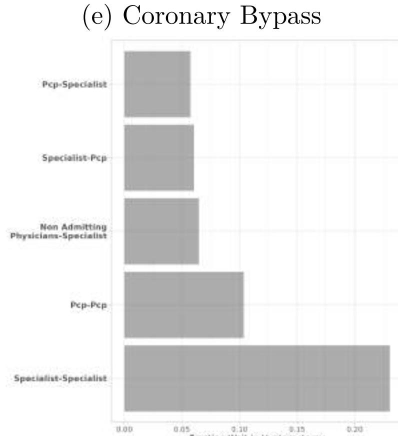
(e) Coronary Bypass



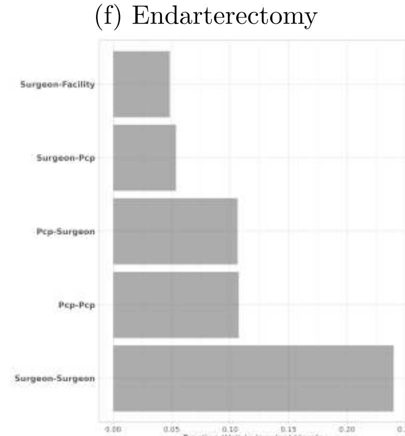
(f) Endarterectomy



(g) Hip Replacement



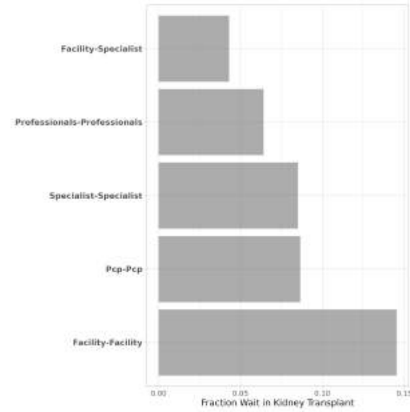
(h) Hysterectomy



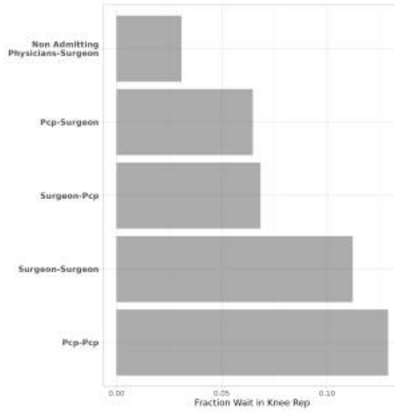
(i) Inguinal Hernia

Figure C19. Components of Wait Times by Surgery

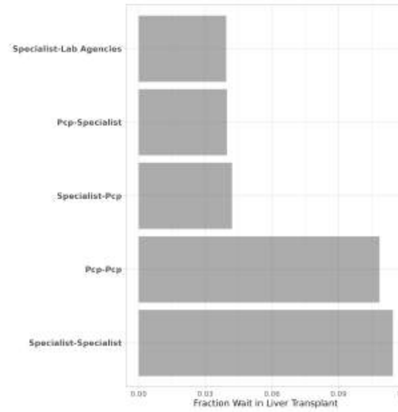
Notes:.



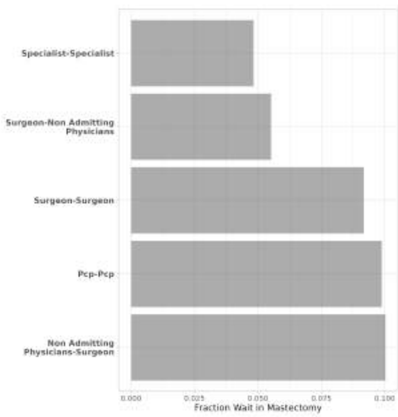
(a) Kidney Transplant



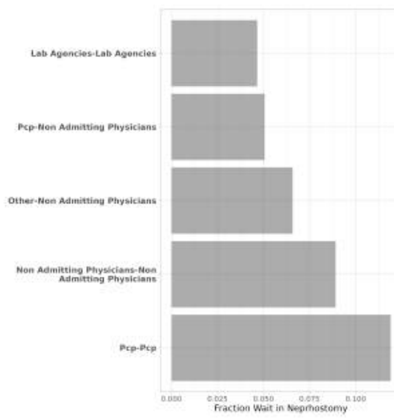
(b) Knee Replacement



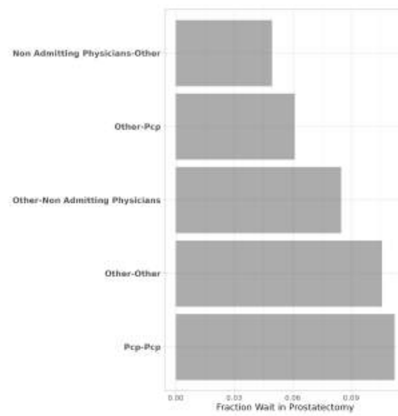
(c) Liver Transplant



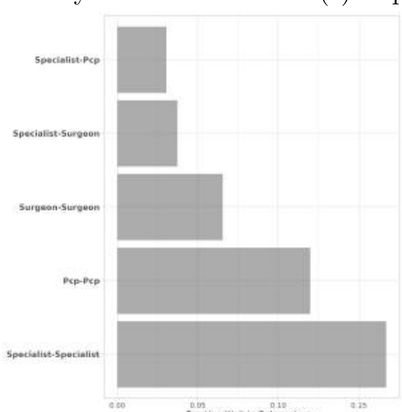
(d) Mastectomy



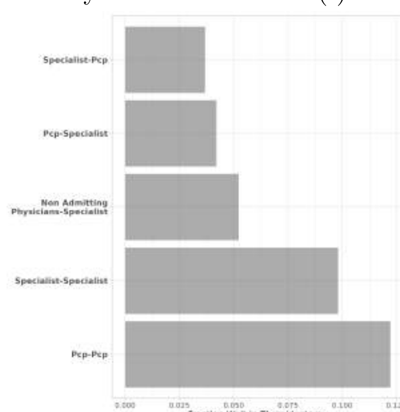
(e) Nephrostomy



(f) Prostatectomy



(g) Splenectomy



(h) Thyroidectomy

Figure C20. Components of Wait Times by Surgery (contd.)

Notes:.

Table C5. Wait Times Components and Insurance Plan Design

Dependent Variables:	Ct Visits	Gap Bw. Visits	Ct Visits	Gap Bw. Visits	Ct Visits	Gap Bw. Visits	Ct Visits	Gap Bw. Visits
			OLS				IV	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
HMO	-0.7601***	0.7834***	-1.421***	0.8409***	-0.0144	1.026***	-0.6680***	1.070***
	(0.1190)	(0.0757)	(0.1323)	(0.0740)	(0.2109)	(0.1339)	(0.1905)	(0.1249)
EPO	-0.5306***	0.1181	-1.410***	0.1947	0.3862	-0.4091	-0.5921	-0.3426
	(0.1866)	(0.1631)	(0.2139)	(0.1743)	(0.7683)	(0.4029)	(0.7299)	(0.3863)
POS	-0.6958***	-0.0522	-1.123***	-0.0150	-0.9028***	0.0011	-0.9114***	0.0017
	(0.0949)	(0.0452)	(0.1245)	(0.0547)	(0.1336)	(0.0957)	(0.1370)	(0.0952)
POScap	0.2333	-0.3564*	-0.5213	-0.2907	3.642***	-1.521***	0.4681	-1.306***
	(0.3703)	(0.1842)	(0.3768)	(0.1832)	(0.9363)	(0.5352)	(0.8540)	(0.4938)
CDHP	-0.7690***	0.2525**	-0.0478	0.1897*	-1.287***	0.1991	-0.1347	0.1207
	(0.1372)	(0.1086)	(0.1581)	(0.1020)	(0.2489)	(0.2063)	(0.2422)	(0.1870)
HDHP	-1.007***	0.1087	0.3182	-0.0067	-0.0487	-0.4383	1.570***	-0.5484*
	(0.2094)	(0.1380)	(0.2049)	(0.1411)	(0.5868)	(0.2743)	(0.5890)	(0.2870)
Comprehensive	-0.3000	0.3412***	0.1483	0.3022***	-0.9962***	0.4983**	-0.3899	0.4570**
	(0.2481)	(0.1156)	(0.2332)	(0.1117)	(0.3734)	(0.2071)	(0.3537)	(0.2014)
Share Out of Pocket			-6.561***	0.5712*			-7.770***	0.5285
			(0.5711)	(0.2910)			(0.7232)	(0.3723)
Fixed-effects								
surg-surg_year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
surg-surg_month	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics								
Observations	844,801	844,801	844,801	844,801	844,801	844,801	844,801	844,801
R^2	0.24929	0.10826	0.25008	0.10828	0.24897	0.10815	0.24981	0.10817
F-test (1st stage), Comprehensive					100,105.3	100,105.3	87,742.6	87,742.6
F-test (1st stage), EPO					26,317.0	26,317.0	23,027.5	23,027.5
F-test (1st stage), HMO					69,116.5	69,116.5	60,571.0	60,571.0
F-test (1st stage), POS					118,509.1	118,509.1	103,723.6	103,723.6
F-test (1st stage), POScap					30,312.5	30,312.5	26,537.1	26,537.1
F-test (1st stage), CDHP					71,521.3	71,521.3	62,602.7	62,602.7
F-test (1st stage), HDHP					23,719.4	23,719.4	20,780.6	20,780.6
F-test (1st stage), Share Out of Pocket							182,001.7	182,001.7
Healthplan	No	No	No	No	No	No	No	No

Clustered (surg-surg_year) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Wait times with detection threshold at $\bar{p} = 0.9$. All dependent variables are in levels: either days or number of visits or procedures. Variables are in levels. Plan types are instrumented by share of plan types at the contributor – year level, and OOP share is intrumented by average OOP share at the contributor – year level.

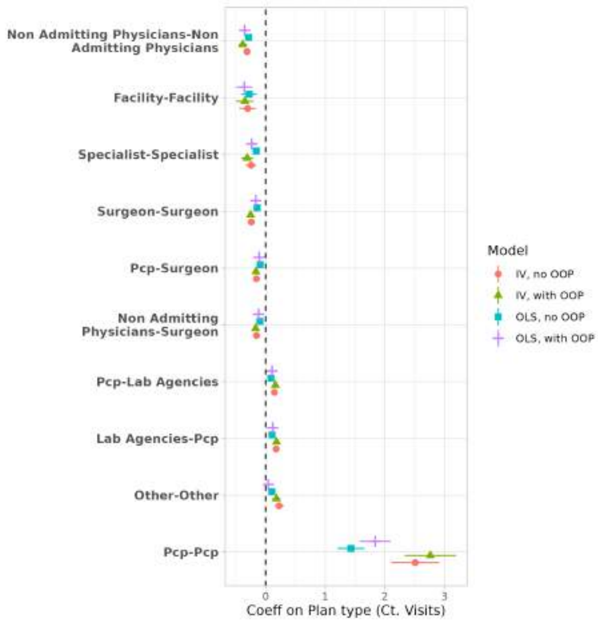
Table C6. Laboratory visits, Preoperative Tests and Plan Design

Dependent Variables:	Ct Lab. Visits	Ct Preop. Tests	Ct Lab. Visits	Ct Preop. Tests	Ct Lab. Visits	Ct Preop. Tests	Ct Lab. Visits	Ct Preop. Tests
	OLS				IV			
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
HMO	0.1900***	0.2636***	0.1300***	0.0992	0.2045***	-0.2433	0.1459***	-0.4287**
	(0.0257)	(0.0946)	(0.0242)	(0.1002)	(0.0333)	(0.1685)	(0.0322)	(0.1766)
EPO	0.1064**	0.1304	0.0266	-0.0884	-0.7197***	-1.388***	-0.8074***	-1.665***
	(0.0514)	(0.1822)	(0.0491)	(0.1842)	(0.1797)	(0.4702)	(0.1849)	(0.4850)
POS	0.2776***	-0.5806***	0.2389***	-0.6869***	0.1843***	-0.9457***	0.1836***	-0.9482***
	(0.0313)	(0.0668)	(0.0280)	(0.0665)	(0.0389)	(0.1198)	(0.0393)	(0.1158)
POScap	0.9574***	0.0575	0.8889***	-0.1303	3.423***	-0.2545	3.138***	-1.155
	(0.1015)	(0.2489)	(0.0985)	(0.2471)	(0.3866)	(0.6761)	(0.3656)	(0.7171)
CDHP	0.0400	-0.8851***	0.1055***	-0.7056***	0.1558***	-2.465***	0.2592***	-2.138***
	(0.0306)	(0.0930)	(0.0369)	(0.0915)	(0.0571)	(0.1964)	(0.0648)	(0.1988)
HDHP	-0.1591***	-0.4175***	-0.0388	-0.0878	-0.5444***	-0.2122	-0.3992***	0.2470
	(0.0431)	(0.1323)	(0.0444)	(0.1443)	(0.1288)	(0.4568)	(0.1204)	(0.4513)
Comprehensive	0.0206	0.3409**	0.0613	0.4524***	0.0358	1.249***	0.0901	1.421***
	(0.0510)	(0.1333)	(0.0529)	(0.1422)	(0.0944)	(0.2033)	(0.1001)	(0.2171)
Share Out of Pocket			-0.5960***	-1.633***			-0.6969***	-2.204***
			(0.0801)	(0.2364)			(0.0904)	(0.2521)
Fixed-effects								
surg-surg_year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
surg-surg_month	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics								
Observations	844,801	844,801	844,801	844,801	844,801	844,801	844,801	844,801
R^2	0.06650	0.18522	0.06670	0.18534	0.06462	0.18434	0.06505	0.18450
F-test (1st stage), Comprehensive					100,105.3	100,105.3	87,742.6	87,742.6
F-test (1st stage), EPO					26,317.0	26,317.0	23,027.5	23,027.5
F-test (1st stage), HMO					69,116.5	69,116.5	60,571.0	60,571.0
F-test (1st stage), POS					118,509.1	118,509.1	103,723.6	103,723.6
F-test (1st stage), POScap					30,312.5	30,312.5	26,537.1	26,537.1
F-test (1st stage), CDHP					71,521.3	71,521.3	62,602.7	62,602.7
F-test (1st stage), HDHP					23,719.4	23,719.4	20,780.6	20,780.6
F-test (1st stage), Share Out of Pocket							182,001.7	182,001.7
Healthplan	No	No	No	No	No	No	No	No

Clustered (surg-surg_year) standard-errors in parentheses

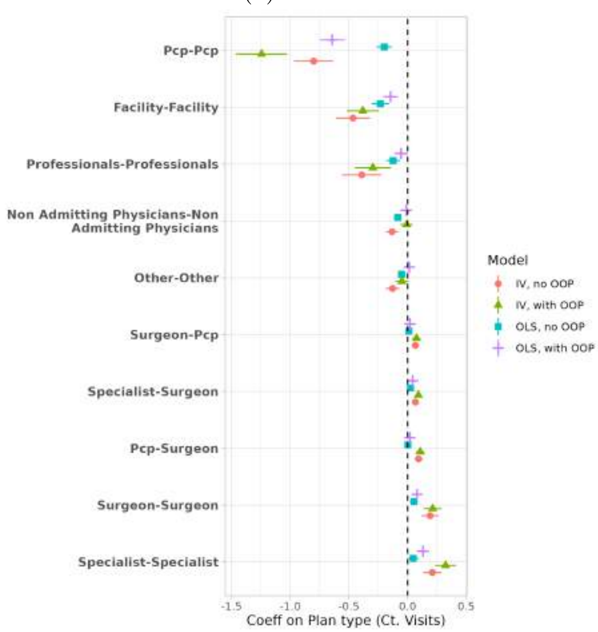
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Wait times with detection threshold at $\bar{p} = 0.9$. Variables are in levels. Plan types are instrumented by share of plan types at the contributor – year level, and OOP share is instrumented by average OOP share at the contributor – year level.

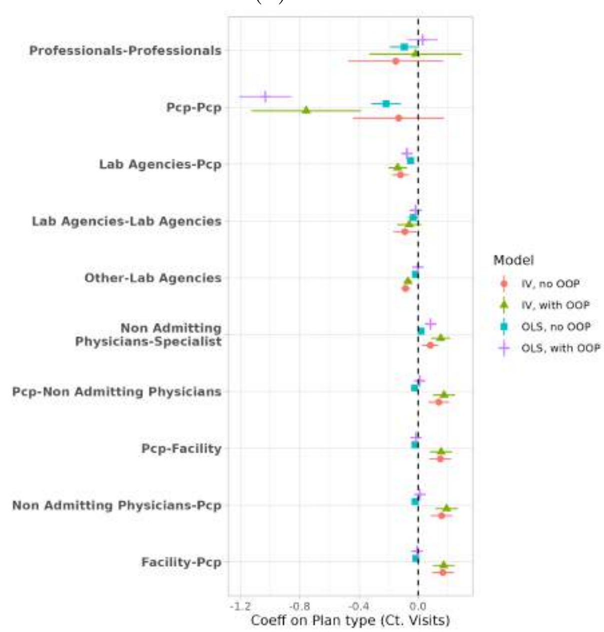


(a) HMO

(b) POS



(c) CDHP



(d) HDHP

Figure C21. Insurance Design and Components of Wait Times (Extensive Margin Only)

Notes: Each Panel shows OLS and IV coefficients and 95% confidence intervals for an insurance plan type relative to PPO, estimated jointly. Specifications either include or exclude controls for plans' OOP share (instrumented in the case of IV). Sample excludes Healthplan contributors. Standard errors are clustered surgery-year .

C.3 Wait Times and Outcomes

C.3.1 Medical Outcomes: Additional Regression Results

Table C7. Descriptive Statistics and Balance, Wait Times and Instrument

Surgery	First Qt.	Fourth Qt.	Diff.	First Qt.	Fourth Qt.	Diff.
	Panel A: Age					
	Wait Time			Instrument		
Cataract	57.40	58.50	0.15	57.40	58.10	0.09
Cholecystectomy	45.00	46.10	0.07	44.90	45.80	0.06
Colectomy	53.30	50.80	-0.21	52.20	51.60	-0.05
Coronary Bypass	56.30	57.40	0.16	56.80	57.00	0.02
Endarterectomy	58.10	59.30	0.22	58.70	58.70	-0.00
Hip Rep	55.30	55.90	0.07	55.40	55.50	0.01
Hysterectomy	45.40	44.30	-0.13	44.90	44.70	-0.02
Inguinal Hernia	45.50	45.40	-0.01	44.80	45.40	0.03
Kidney Transplant	45.10	48.90	0.22	47.10	47.90	0.05
Knee Rep	57.00	56.90	-0.01	56.70	56.90	0.03
Liver Transplant	51.10	52.80	0.11	52.80	50.90	-0.13
Mastectomy	52.70	52.10	-0.07	52.50	52.30	-0.02
Prostatectomy	56.80	58.00	0.21	57.30	57.20	-0.01
Thyroidectomy	46.90	48.60	0.13	47.80	47.90	0.01
	Panel B: Charlson Score					
	Wait Time			Instrument		
Cataract	0.60	1.30	0.38	0.80	0.90	0.05
Cholecystectomy	0.70	1.10	0.23	0.70	0.90	0.11
Colectomy	3.20	1.90	-0.34	2.50	2.20	-0.09
Coronary Bypass	0.80	1.70	0.46	1.10	1.30	0.06
Endarterectomy	0.80	1.50	0.41	1.00	1.20	0.07
Hip Rep	0.60	0.80	0.07	0.70	0.70	0.02
Hysterectomy	0.70	0.60	-0.04	0.60	0.60	-0.01
Inguinal Hernia	0.40	0.60	0.09	0.40	0.50	0.07
Kidney Transplant	2.60	2.90	0.12	2.70	2.70	-0.00
Knee Rep	0.60	0.80	0.12	0.70	0.70	0.06
Liver Transplant	7.70	8.20	0.17	7.90	8.00	0.02
Mastectomy	3.20	4.10	0.28	3.40	3.70	0.09
Prostatectomy	2.50	2.60	0.04	2.60	2.60	0.01
Thyroidectomy	2.30	1.40	-0.31	2.00	1.90	-0.02

Notes: Wait times with detection threshold at $\bar{p} = 0.9$. Sample only includes patients from employer contributors. *First Qt.* and *Fourth Qt.* columns stand for the mean of the covariate considered (age in Panel A, Charlson index in Panel B) for patients in the first or fourth quartile of wait time and of the instrument. *Diff.* columns contain the standardized mean difference between the fourth quartile and the first quartile means.

Table C8. Impact of Wait Times on Inpatient Payments, Alternative Specifications

Dependent Variable:	Inp. Pay ($)							
	OLS				IV			
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Wait Time 0.9 (days)	2.987***	3.035***	3.019***	2.047***	11.09***	9.794***	9.745***	7.251***
	(0.1804)	(0.1821)	(0.1818)	(0.1793)	(1.935)	(1.774)	(1.684)	(1.751)
Patient Controls	No	No	No	Yes	No	No	No	Yes
Fixed-effects								
Surgery	Yes				Yes			
Surgery-Plan		Yes	Yes	Yes		Yes	Yes	Yes
Surgery-Year			Yes	Yes			Yes	Yes
Surgery-Month			Yes	Yes			Yes	Yes
Fit statistics								
Observations	762,466	762,466	762,466	762,466	762,466	762,466	762,466	762,466
R^2	0.02813	0.05049	0.05256	0.06173	0.02422	0.04782	0.04992	0.06019
F-test (1st stage), Wait Time 0.9 (days)					11,496.7	12,426.1	13,515.2	12,609.6

Heteroskedasticity-robust standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Wait times with detection threshold at $\bar{p} = 0.9$. Sample excludes Appendectomy, Nephrostomy, and Splenectomy for which the detection models perform poorly, and only includes patients from employer contributors. Dollars are deflated to 2010 dollars.

Table C9. Impact of Wait Times on Opioids Prescription, Alternative Specifications

Dependent Variable:	Opioids (days supp.)							
	OLS				IV			
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Wait Time 0.9 (days)	0.0163***	0.0163***	0.0163***	0.0135***	0.0432***	0.0422***	0.0450***	0.0397***
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0026)	(0.0025)	(0.0024)	(0.0025)
Patient Controls	No	No	No	Yes	No	No	No	Yes
Fixed-effects								
Surgery	Yes				Yes			
Surgery-Plan		Yes	Yes	Yes		Yes	Yes	Yes
Surgery-Year			Yes	Yes			Yes	Yes
Surgery-Month			Yes	Yes			Yes	Yes
Fit statistics								
Observations	762,466	762,466	762,466	762,466	762,466	762,466	762,466	762,466
R^2	0.07474	0.08907	0.08968	0.10552	0.06419	0.07944	0.07784	0.09593
F-test (1st stage), Wait Time 0.9 (days)					11,496.7	12,426.1	13,515.2	12,609.6

Heteroskedasticity-robust standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Wait times with detection threshold at $\bar{p} = 0.9$. Sample excludes Appendectomy, Nephrostomy, and Splenectomy for which the detection models perform poorly, and only includes patients from employer contributors.

Table C10. Impact of Wait Times on Inpatient Payments, Alternative Specifications, Managed Care Only

Dependent Variable:	Inp. Pay ($)							
	OLS				IV			
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Wait Time 0.9 (days)	2.918***	2.956***	2.929***	1.959***	12.43***	9.676**	9.876**	7.461*
	(0.3490)	(0.3524)	(0.3518)	(0.3540)	(4.519)	(3.996)	(3.852)	(4.032)
Patient Controls	No	No	No	Yes	No	No	No	Yes
Fixed-effects								
Surgery	Yes				Yes			
Surgery-Plan		Yes	Yes	Yes		Yes	Yes	Yes
Surgery-Year			Yes	Yes			Yes	Yes
Surgery-Month			Yes	Yes			Yes	Yes
Fit statistics								
Observations	275,692	275,692	275,692	275,692	275,692	275,692	275,692	275,692
R^2	0.01935	0.03757	0.04023	0.04697	0.01518	0.03553	0.03806	0.04564
F-test (1st stage), Wait Time 0.9 (days)					3,169.1	3,530.8	3,815.5	3,541.2

Heteroskedasticity-robust standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Wait times with detection threshold at $\bar{p} = 0.9$. Sample excludes Appendectomy, Nephrostomy, and Splenectomy for which the detection models perform poorly, and only includes patients from employer contributors. In addition, we exclude PPO, HDHP, CDHP plans to focus only on managed-care plans in which patients have less flexibility in choosing between providers. Dollars are deflated to 2010 dollars.

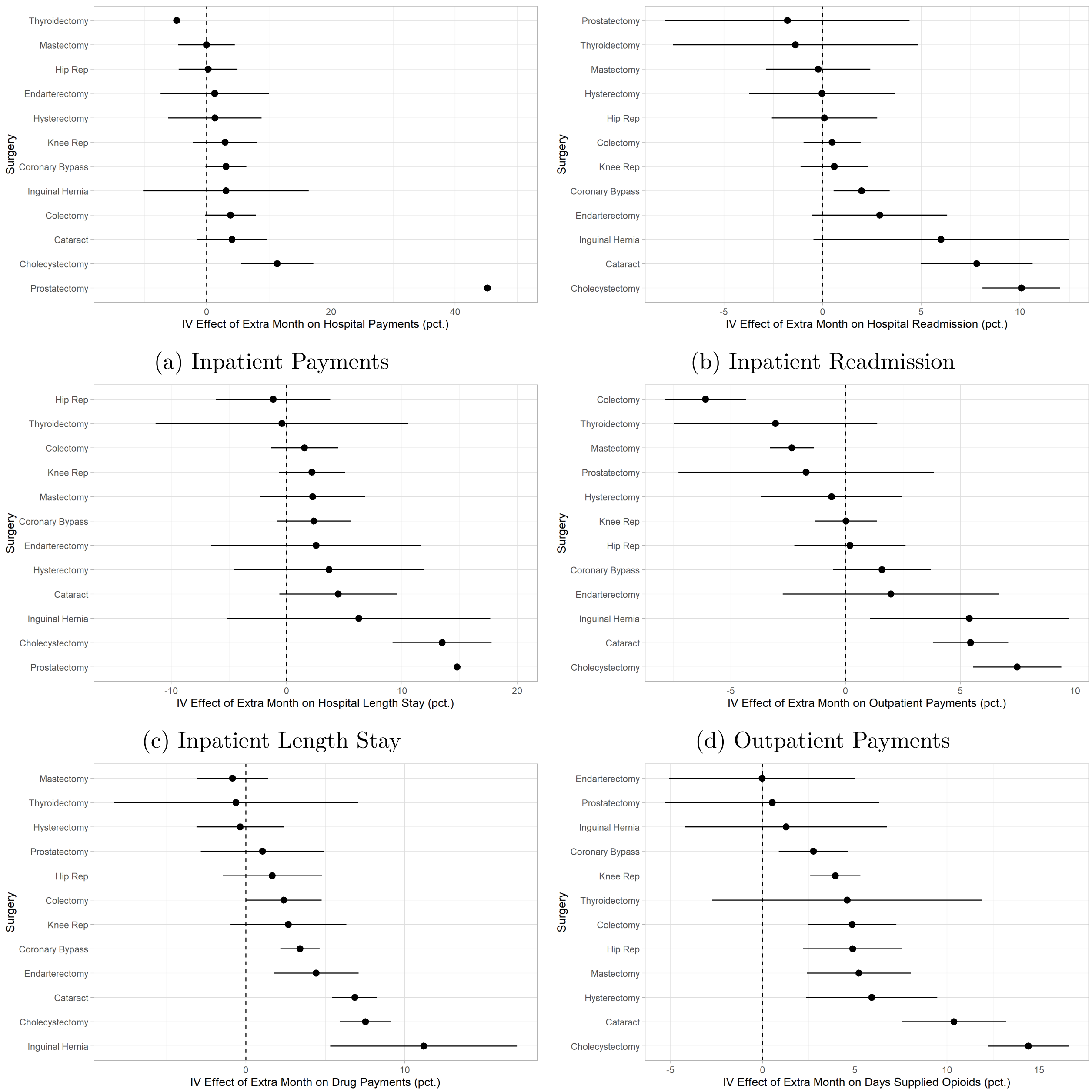
Table C11. Impact of Wait Times on Opioids Prescription, Alternative Specifications, Managed Care Sample

Dependent Variable:	Opioids (days supp.)							
	OLS				IV			
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Wait Time 0.9 (days)	0.0156***	0.0154***	0.0155***	0.0128***	0.0389***	0.0379***	0.0410***	0.0358***
	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0049)	(0.0047)	(0.0046)	(0.0047)
Patient Controls	No	No	No	Yes	No	No	No	Yes
Fixed-effects								
Surgery	Yes				Yes			
Surgery-Plan		Yes	Yes	Yes		Yes	Yes	Yes
Surgery-Year			Yes	Yes			Yes	Yes
Surgery-Month			Yes	Yes			Yes	Yes
Fit statistics								
Observations	275,692	275,692	275,692	275,692	275,692	275,692	275,692	275,692
R^2	0.07439	0.09346	0.09458	0.10988	0.06690	0.08668	0.08583	0.10289
F-test (1st stage), Wait Time 0.9 (days)					3,169.1	3,530.8	3,815.5	3,541.2

Heteroskedasticity-robust standard-errors in parentheses

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Notes: Wait times with detection threshold at $\bar{p} = 0.9$. Sample excludes Appendectomy, Nephrostomy, and Splenectomy for which the detection models perform poorly, and only includes patients from employer contributors. In addition, we exclude PPO, HDHP, CDHP plans to focus only on managed-care plans in which patients have less flexibility in choosing between providers.



(a) Inpatient Payments

(b) Inpatient Readmission

(c) Inpatient Length Stay

(d) Outpatient Payments

(e) Drug Payments

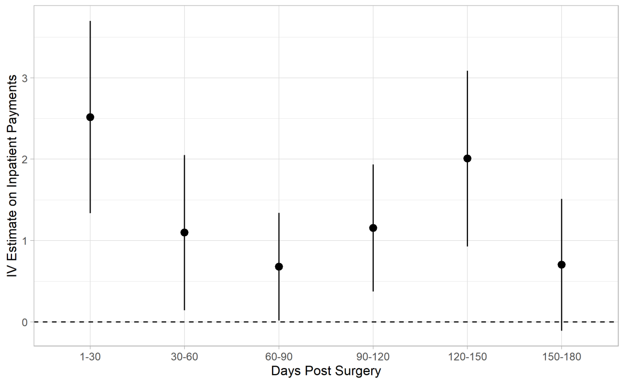
(f) Opioids Prescribed

Figure C22. Causal Effects of Extra Month of Wait by Outcome and Surgery

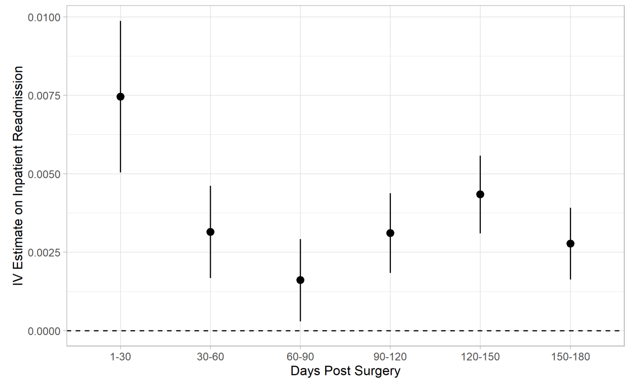
Notes:.

C.3.2 Heterogeneous Effects Analysis

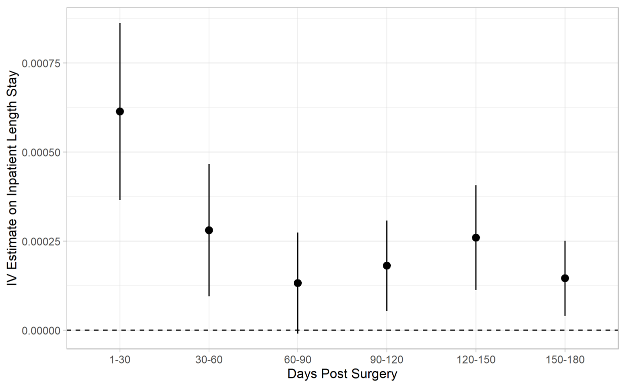
C.3.3 Counterfactuals



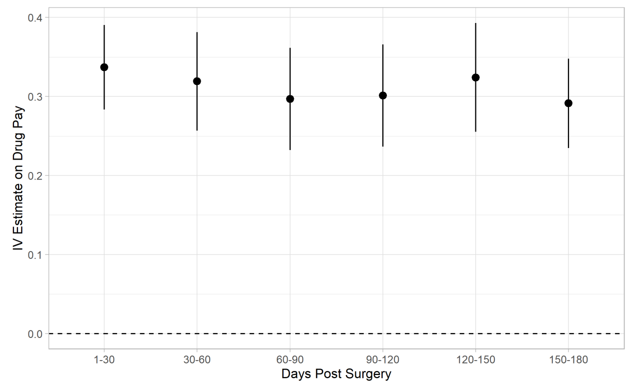
(a) Inpatient Payments



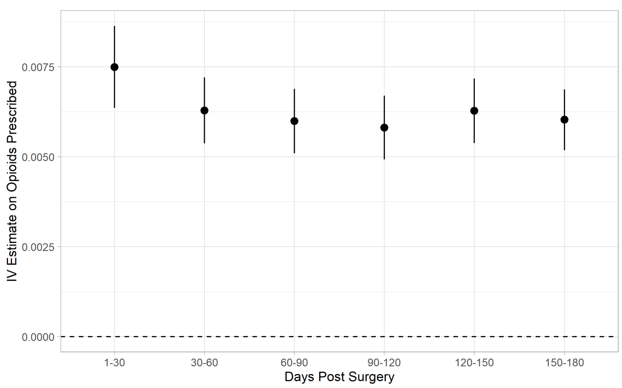
(b) Inpatient Readmission



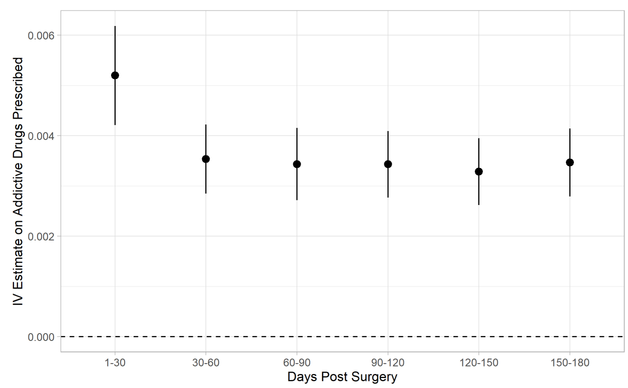
(c) Inpatient Length Stay



(d) Drug Payments



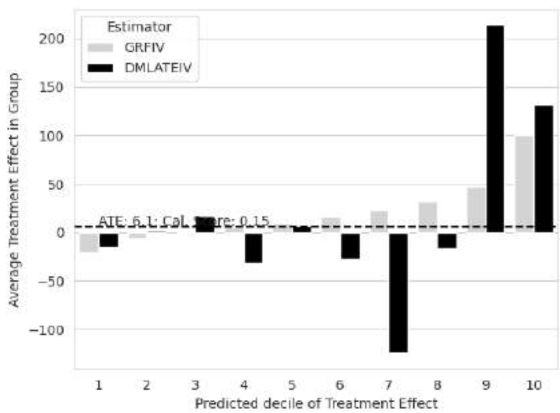
(e) Opioids Prescribed



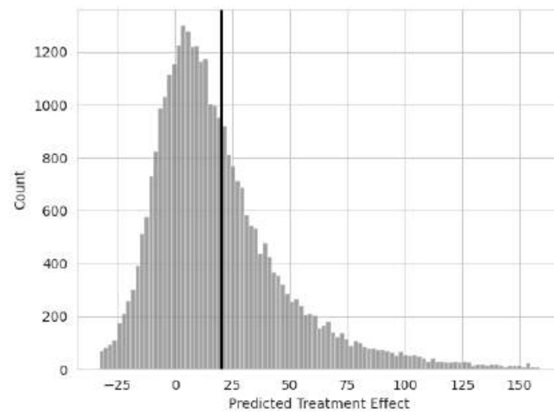
(f) Addictive Drugs Prescribed

Figure C23. Persistence of Adverse Effects of Wait Time, Entire Sample

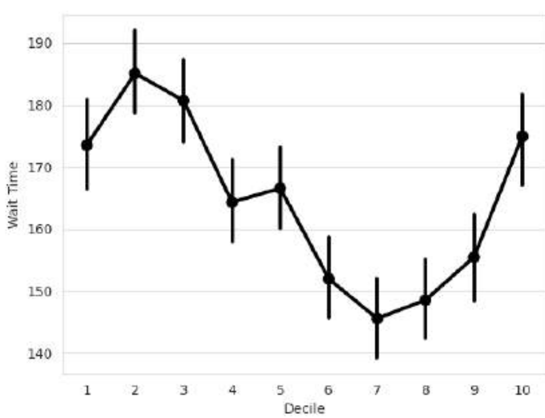
Notes:.



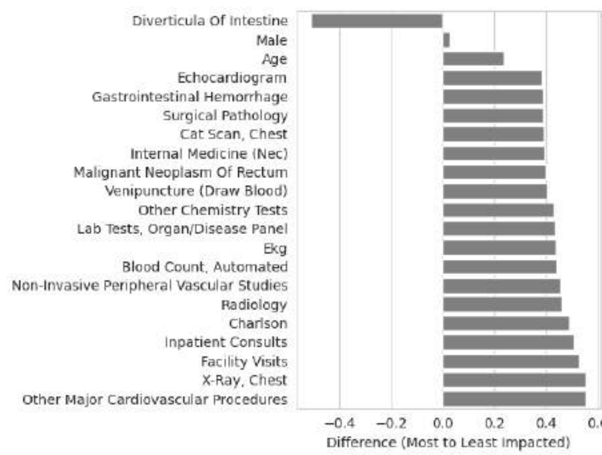
(a) Calibration Plot



(b) Distribution of Effects



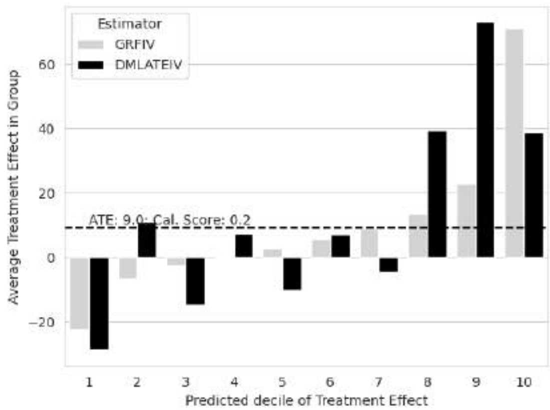
(c) Allocative (In)Efficiency



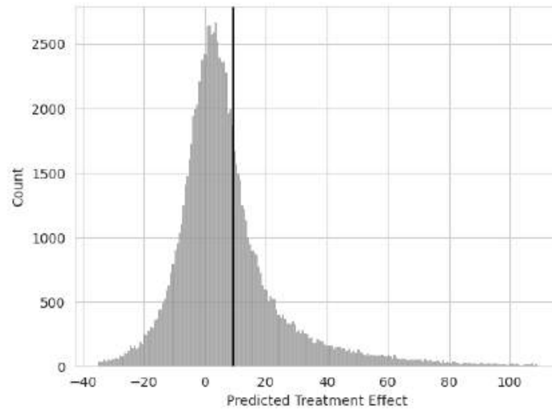
(d) Characteristics of Most Impacted Patients

Figure C24. Heterogeneous Effects from GRF for Colectomy

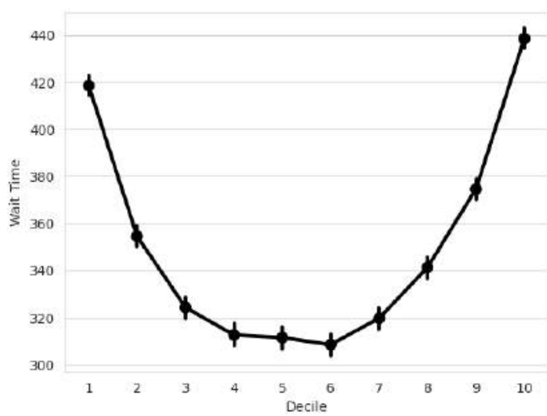
Notes: Panel (a) displays the figure resulting from the calibration procedure in which we rank patients according to deciles of predicted effects from a heterogeneous effects model and then compare the average of the effects ***predicted*** by that model in each decile to the average effect ***estimated*** in each decile by DMLATEIV. The implied calibration score and the full sample average effect estimated by DMLATEIV are also reported on panel (a). Panel (b) shows the distribution and mean of heterogeneous effects. Panel (c) displays average wait times and standard deviations by decile of predicted effects, where the tenth decile contains the patients the most impacted by wait times. Panel (d) shows the difference in standardized patient characteristics between patients in the fourth quartile of effects and patients in the first quartile of effects.



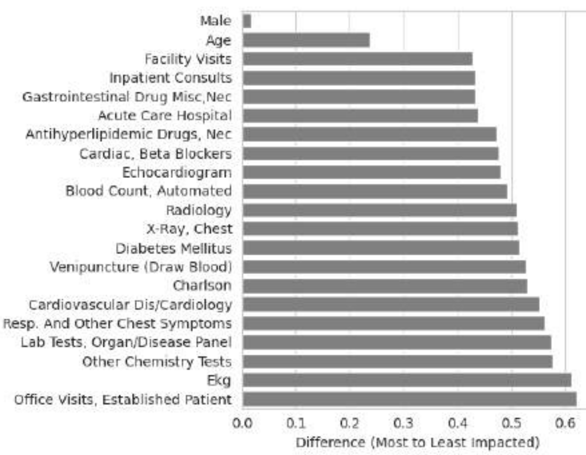
(a) Calibration Plot



(b) Distribution of Effects



(c) Allocative (In)Efficiency



(d) Characteristics of Most Impacted Patients

Figure C25. Heterogeneous Effects from GRF for Knee Replacement Surgery

Notes: Panel (a) displays the figure resulting from the calibration procedure in which we rank patients according to deciles of predicted effects from a heterogeneous effects model and then compare the average of the effects ***predicted*** by that model in each decile to the average effect ***estimated*** in each decile by DMLATEIV. The implied calibration score and the full sample average effect estimated by DMLATEIV are also reported on panel (a). Panel (b) shows the distribution and mean of heterogeneous effects. Panel (c) displays average wait times and standard deviations by decile of predicted effects, where the tenth decile contains the patients the most impacted by wait times. Panel (d) shows the difference in standardized patient characteristics between patients in the fourth quartile of effects and patients in the first quartile of effects.