Treating Doctors: Evidence from a Field Experiment in Surgery *

Hannah Trachtman, Naomi Friedman-Sokuler,
Juan Pablo Atal, Jorge Ale-Chilet, Dina Orkin

March 17, 2025

PRELIMINARY

Abstract

Managing skilled labor is challenging due to the inherent asymmetric information between workers and and management and the importance of workers' professional judgment. Managing healthcare personnel is a prime example where excessive costs are pervasive but directly costs has often proven ineffective or counterproductive. We implement a randomized control field experiment in the largest public hospital in Israel to investigate whether a behavioral intervention may improve the productivity of operating room staff, surgeons and nurses. We find that a cheap, scalable, non-financial intervention, reduces the costs of targeted disposable surgical equipment by 9-11% without affecting clinical outcomes of patients. The effect we identify is fully driven by surgeons who enjoy a very high degree of autonomy vis-á-vis management, while nurses are largely unaffected by our intervention. Finally, we exploit administrative and survey data to examine which design features and team characteristics mediate treatment effects.

Keywords: Management, High skilled labor, Hospital spending, Field experiment

JEL Classification: M54, I10, D91, J24, C93

*Trachtman: Hebrew University of Jerusalem; Friedman-Sokuler: Bar-Ilan University; Atal: University of Pennsylvania; Ale-Chilet: Universidad de los Andes, Chile; Orkin: Sheba Medical Center Hospital

1 Introduction

In labor settings with highly skilled experts, information asymmetries limit the available tools for management. Health care is an important example: managers want physicians to administer cost-effective care, but only physicians know what cost-effective care is. Indeed, financial and even non-financial incentives have resulted in adverse effects on patient care (Alexander, 2020; Silver, 2021; La Forgia, 2023; League, 2023).[1] Research on the management of highly skilled experts performing complex, multi-dimensional tasks is therefore critical.

In this paper, we study whether a non-financial intervention, designed specifically for highly skilled experts, can improve cost-effectiveness in health care. Health care offers an ideal setting to study the management of high-skilled labor: excessive costs are pervasive but providers have limited incentives for cost containment, in large part due precisely to the challenges described above (Cooper et al., 2020; Einav et al., 2021). We ask whether a cheap and scalable non-financial intervention can shift the behavior of medical professionals such that costs fall but the quality of care is preserved. We then examine how the intervention works—which components are essential (or not), which surgeries and items show shifts in usage, and which professionals respond—in pursuit of underlying insights about the effective management of high-skilled professionals.

We focus specifically on the costs of disposable equipment in surgery, which has several advantages. First, surgery is a high-stakes environment where professional expertise and skills are key. Second, economic stakes are sizable: expenditure on medical products, appliances, and equipment account for 15 percent of total EU government expenditure on health, and 1.2 percent of total GDP (Eurostat, 2022).[2] Finally, episodes are well defined, allowing us to observe the medical team in charge, input costs, and outcomes.

We implemented a field experiment in the operating room (OR) wing of the largest hospital in Israel. We designed an intervention that would highlight the costs of disposable items without flagging any particular usage as wasteful. Treated personnel received surgery-specific benchmarking information about typical usage of disposable items, as well as a general encouragement to avoid the preemptive opening of disposable equipment. The benchmarking information provided was explicitly descriptive, without masquerading as any kind of clinical recommendation. It was communicated to treated per-

[1] There are examples outside of health care as well: see for example Bergstresser and Philippon (2006).
[2] Disposable inputs cost our hospital partner in the experiment USD 26.3 million annually.

sonnel via WhatsApp text messages sent at the beginning of each shift, and via periodic messages that provided feedback on completed surgeries. It targeted a wide set of surgeries and items, allowing experts to potentially reduce usage where they see fit, and not where it may go against their professional judgment. Finally, the intervention required minimal information from management, was cheap to implement, and was easily scalable.

The medical professionals responsible for the disposable equipment in the OR are surgeons and circulating nurses. Surgeons are the ultimate decision-makers within the OR, are not subordinate to the OR wing manager (who oversees the budget) and enjoy a very high degree of autonomy vis-à-vis their own direct managers, the surgical department heads. Circulating nurses are responsible for the flow of equipment into the OR and for opening items during surgery. The circulating nurse is subordinate to the surgeon in the OR and is directly managed by the OR Wing. By administering the intervention not only to surgeons but also to nurses, we can explore whether the effectiveness of such an intervention depends the position of the expert within the organization. We thus randomly assigned half of surgeons and half of circulating nurses to our intervention, generating four types of teams. The intervention was implemented over nine months and 9,465 scheduled surgeries.

We find that the intervention on surgeons has statistically significant and economically large effects on surgery-level costs. Although treating the surgeon does not cause discernible effects on the number of inputs used on average, we find large decreases for expensive targeted inputs. As a result, our intervention decreases expenditures on targeted inputs by 9-11% per surgery, without producing any detectable effects on patients' post-operative ICU referral, imaging, or time in surgery. The effect of treating the nurse is not significantly different from zero, neither for surgery costs nor for item usage.

We argue that the intervention shifted surgeon behavior by making the costs of disposable items salient. Several pieces of evidence, exploiting plausible exogenous variation in the informational content of the intervention, go *against* the idea that item-specific benchmarking information was essential. Moreover, the dynamics of the treatment effect demonstrate only partial persistence. The effect of treating the surgeon is robust to temporary disruptions, but it gradually falls as the last received message recedes in time, and after the end of the intervention. This evidence is consistent with a story in which the intervention worked not because it permanently shifted beliefs, but because it temporarily made costs salient. It aligns with other evaluations of nudges that find effects to be large

at the beginning when the intervention is novel, and then to wane over time, especially once the intervention is removed.

Finally, we argue that the success of such an intervention depends crucially on the position of the expert within the organization. We find no impacts of our intervention on circulating nurses who, as opposed to surgeons, do not have final say over equipment use, and who are embedded in a traditional management structure. In a baseline survey, we collected data on surgeon and nurse attitudes toward cost-containment and decision-making power. It seems that the average circulating nurse, already operating under traditional management structures, is perhaps not the kind of expert that requires or benefits from such an intervention.

We contribute, first, to a small but important literature on the management of expert workers engaging in non-routine tasks. Most evidence on personnel management comes from settings with routine tasks and low-skilled workers, with a few notable exceptions. Our paper is closely related to the work of Gosnell et al. (2020), who test a set of management practices in a field experiment with professional airline captains, and find strong and persistent effects of monitoring as well as personalized targets on performance. Englmaier et al. (2024) also study skilled workers in non-routine tasks, and find positive effects of financial incentives on team performance. Both of these settings are fundamentally different from surgery in two important ways. First, it is actually possible to measure efficiency at the level of the individual episode. Second, the stakes for unintended consequences of interventions are far lower.[3] Our intervention is designed for settings where managers must trust experts to identify waste.

Second, we contribute to an extensive literature on policies that promote cost-effective health care among medical personnel. A large literature evaluates financial incentives for cost-effective care in a wide variety of contexts.[4] In some cases, financial incentives have worked well (Eliason et al., 2022); in others, they have produced unintended consequences (Alexander, 2020; La Forgia, 2023).[5] There is also a nascent literature on standard cost-containment policies like auditing, claim denial, and anti-fraud (Shi, 2024; Leder-

[3] In Gosnell et al. (2020), all targeted behaviors were well within the relevant safety standards.

[4] A series of lab experiments evaluate common payment schemes such as fee-for-service and capitation (Hennig-Schmidt et al., 2011; Brosig-Koch et al., 2016; Brosig-Koch et al., 2017). Mohanan et al. (2021) investigate health provider behavior under contracts linked to health inputs versus health outputs.

[5] Alexander (2020) finds that bonuses for reducing total hospital costs of admitted Medicare patients only produced patient selection, without cost reduction. La Forgia (2023) shows that when medical practices hire management companies that prioritize financial management, they provide less appropriate care in childbirth.

Luis, 2023; Howard and McCarthy, 2021; Eliason et al., 2021). These policies are often effective, imposing large consequences on providers who do not comply with pre-defined rules about what constitutes appropriate care or expenditure. However, such policies are not relevant for many contexts, including disposable item use in surgery, where standards of appropriate expenditure are harder to define. Other papers evaluate "softer" non-financial interventions or nudges. Malhotra et al. (2016) look at the use of non-generic drugs, a setting where instances of waste are observable, and find that defaults effectively promote the use of generics. Doyle et al. (2019) evaluate the use computer-decision-support (CDS) for imaging, which, like our intervention, complements and relies on expert judgment. They find that CDS reduces imaging orders, but the implications for cost-effective care are not clear.[6]

Finally, we build on the very large literature about inefficiencies in health care provision more generally. In particular, we add to an emerging literature that attempts to identify particular case studies of waste in health care and propose solutions (see for example Einav et al. (2021); Cooper et al. (2020)). We highlight a new setting where there might be inefficiencies—disposable input use in surgery—and show that a simple intervention can shift usage without reducing the quality of care.

The remainder of the paper proceeds as follows. In Section 2 we describe the context, data, and experiment design. In Section 3 we present our main results on the effect of the intervention and explore mechanisms. In Section 4 we conclude.

2 Experiment Design

2.1 Context and Data

The experiment was designed with and implemented within the OR wing of Sheba Tel HaShomer Health City (Sheba), Israel's leading and largest hospital.[7] The OR wing is a service department within the hospital that manages all physical and personnel inputs

[6]Our paper contributes to a wider movement that applies ideas from behavioral economics to shift provider behavior and improve care (Emanuel et al., 2015; Navathe et al., 2016; Navathe and Emanuel, 2016; Patel et al., 2023). This movement has also inspired many nudges targeting clinical outcomes: see for example Zureich and Singh (2024); Sacarny et al. (2023, 2022); Gould et al. (2017); Kolstad (2013).

[7]Sheba hospital is a public hospital whose funding as well as its labor relations are governed by the state and its collective agreements.

related to surgery, with the exception of the surgeons themselves that belong to surgical departments. The wing comprises 32 operating rooms serving all 18 surgical departments in the hospital, performing over 400 types of surgical procedures in about 30,000 surgeries annually.[8] OR wing staff includes all OR nurses and anesthesiologists as well as administrative and logistical staff. All surgical equipment is budgeted and managed by the OR wing, including more than 10,000 types of disposable medical items that cost USD 25 million yearly.

The study was made possible by the introduction, in early 2019, of a novel inventory system tracking the usage of disposable equipment at the individual surgery level, linked both to patients and personnel. When this data became available, the costs of disposable equipment were found to be considerably higher than the OR wing management expected, motivating the initiation of this research. For this reason, the focus of our intervention is the use of disposable equipment in surgery, measured by costs or by units consumed.

The main unit of observation in this study is a *surgery*, an event in which a *procedure* (i.e., knee replacement) is performed by a *team* on a *patient*. The vast majority of surgical procedures require disposable equipment, which refers to both single-use and multiple-use items. Multiple-use items are often more expensive and complex (e.g., harmonic scalpels), and can be used and sterilized up to five times before they need to be discarded. Single-use items include things like blades, drapes, and wound dressings. Scheduled surgeries are logged into the OR logistical system in advance, and OR wing staff prepares a cart containing a pre-specified procedure-specific list of disposable equipment.[9] At the end of the surgery, opened disposable equipment is discarded, whether it was used or not. Multiple-use equipment, as well as unopened disposable items, are returned to stock after sterilization.

While the surgical team comprises of a variety of medical personnel—surgeons, surgical interns, scrub nurse, circulating nurse, anesthesia staff and others—we worked together with the OR management team to identify the personnel relevant for the use of disposable equipment. Two teams members were identified: *surgeons*, the principal decision-

[8]The bulk of scheduled surgeries occur during the main operating hours, 8am to 3pm, and emergency surgeries occur throughout the day and night. In addition, public hospitals are allowed to choose surgeries to perform in an afternoon session 3pm-7pm, to shorten waiting times for surgeries and supplement their income. However, they can perform at most 40% of every procedure type in the afternoon. All surgeries, regardless of the session, are paid by the national HMOs according to the Ministry of Health's price list.

[9]See Appendix Figure A1.

makers in the operating room; and the *circulating nurses*, who oversee the inventory of surgical instruments, supplies, and equipment needed for the procedure. Both surgeons and nurses clearly engage in 'skilled labor': they possess strong professional identities and a sense of social obligation and organizational mission, they embody significant human capital, and they operate in a high-stakes environment (Gosnell et al., 2020).[10] At the same time they differ in their position in the organizational structure. First, surgeons have considerably more decision making power over equipment use and ensuing costs than do nurses, who like all other personnel in the OR, are subordinate to the lead surgeon. Second, surgeons are managed by their respective departments, and the costs of disposable equipment do not directly affect them or their managers. Nurses, on the other hand, are managed by the Head Nurse who is part of OR wing management. Since the OR wing management is responsible for the budget and cares about costs, it can (and does) employ traditional management practices to align nurse incentives with their own.

In our conversations with the OR wing management prior to the study, two hypotheses were raised about potential sources of equipment overuse. The first is misuse: some of the equipment that is brought to the OR may not be necessary for the surgery, or it may be easily substituted with a cheaper alternative. The second is excessive precautionary opening: once an item has been brought to the OR, it might be opened in advance unnecessarily. Both surgeons and circulating nurses potentially contribute to these sources of overuse: surgeons are the final decision-makers, but nurses physically bring and open equipment. Ultimately, we cannot identify instances of waste, nor can we be sure that there is waste at all, and we will design our intervention accordingly.

The core dataset used in the design of the intervention as well as the analysis is derived from the OR inventory system described above. We augment this data set with patient level data retrieved from the hospital's clinical management system. This dataset contains patients' American Society of Anesthesiologists (ASA) physical status score, a benchmark for assessing the fitness of patients before surgery; and post-op outcomes such as time in surgery, imaging, and ICU referral.

Additionally, we conducted a baseline survey among participating personnel, in which we collected demographic information as well as questions on four topics: observed

[10]Both surgeons and nurses are directly employed by the hospital on the basis of collective agreements, that are profession-specific but common across all public hospitals. The basic employment contract for both surgeons and nurses includes morning shifts (7am to 3pm) and on-call shifts (emergencies). In addition, surgeons and nurses can opt to work in the afternoon session (3pm-7pm), where they receive supplemental pay per surgery.

patterns of input use in the OR; beliefs regarding own competency; attitudes regarding surgeon-nurse teamwork; and attitudes regarding role of cost considerations in health care provision. Finally, we collected data from the messaging system, which we describe below.

2.2 Intervention Design

In a context with highly skilled labor and managers who are fundamentally unable to observe instances of waste, designing an intervention that shifts behavior on the desired margins is not trivial. Over many discussions with the OR wing management, we devised an intervention, tailored specifically for this context in three ways. First, the intervention provides benchmarking information on historical item usage by procedure type. The information provided is a description of historical usage, and is explicitly not a clinical recommendation. Moreover, the language used is soft. Our intention was to respect expertise and limit the risk of unintended consequences.[11] Second, the intervention is designed to attract the attention of busy surgeons and nurses. The benchmarking information we provide is timely, customized, novel, and potentially activates motivations related to social norms. It is delivered in three ways: as a prompt before the relevant surgery, as a database that can be accessed at any moment, and as feedback several weeks after. Finally, the intervention targets a wide set of surgeries and items, restricted only by statistical criteria. Since the same information asymmetries that inhibit easy management also preclude effective targeting of surgeries, items, or personnel, this strategy allows experts to make changes as they see fit. In what follows, we first describe the construction of the benchmarking information, and then detail its delivery.

Candidate Items and Benchmarks

We restricted the focus of the intervention to a set of item-procedure combinations that are sufficiently common and exhibit sufficient variation. Specifically, we used two years of data to classify as a *candidate item* any item-procedure combination that met the following conditions: (1) the procedure was performed at least 30 times in the 2 year period; (2) the item was used in at least 50% of surgeries within the procedure type; (3) the items

[11] An important feature of our setting, differing from previous management interventions (Bloom et al., 2020; Gosnell et al., 2020), is that management cannot set universal item usage targets, even by procedure, as it is neither feasible nor clinically desirable.

were used in surgeries with at least two distinct surgeons and two distinct nurses; (4) the standard deviation of usage within procedures was above the 25 percentile. The process yielded a total of 4,826 candidate items in 203 procedures. The number of candidate items among these procedures ranges from 4 to 55, and the median procedure has 21 candidate items. Among the candidate items we classify as "inexpensive" items with a unit price less than USD 2. Inexpensive items constitute about half of the candidate items.[12]

The *benchmark amount*, as approved by the medical professionals on the research team, is the median quantity of a candidate item used in that procedure in the 24 months preceding the experiment. We introduced the benchmark amount on the day before the experiment began. Treated personnel received an opening WhatsApp message, explaining the benchmark as the median and asserting that it is not a clinical recommendation (see full text in Appendix A2).

Intervention Components

The benchmarking information was delivered in the following three ways. Images of all messages can be found in Appendix Figure A2, and all translated message texts can be found in Appendix A2.

1. **Pre-Shift Messages:** At the beginning of each morning and afternoon shift, treated personnel who are scheduled for surgeries in that shift received a personally customized WhatsApp text message to their mobile phone. The message comprised a general prompt as well as benchmark amounts for three randomly chosen items in the upcoming procedure. For on-call shift, where there are no scheduled procedures, the message contained the general prompt and a referral to the website (see (2) below).

   When listing the three benchmark amounts, we explicitly reference the median each time to make it very clear to the participants that the benchmark is not a clinical guideline but a statistic of actual usage by themselves and their colleagues.

   When we randomly select three candidate items, we ensure that no more than one is inexpensive in order to weigh our treatment toward higher value items.

[12] We also designated as "control candidate items" half of the items that fall in the bottom price quartile. These items never received messages and were intended as a comparison group to the other half of the bottom quartile who did receive messages.

2. **Website:** All treated personnel received a link to a website with a searchable table of medians for all candidate items by procedure. They could access this link throughout experiment. This link was also included in messages for on-call shifts.[13]

3. **Feedback:** Treated surgeons and nurses received a bi-weekly message that provided feedback on the three candidate inputs that were referenced in a message prior to a surgery conducted over the last 14 days. The message compared the actual amounts that were used to the benchmark amounts. When actual usage was at or below the benchmark, a green check mark appeared alongside the item. When actual usage exceeded the benchmark, a red X appeared alongside the item.

2.3 Sample and Descriptive Statistics

In this section, we first describe our sample frame. We then use this sample to provide descriptive statistics from the pre-period, first on surgery costs and item usage, and second on personnel. Finally, we report the numbers of surgeries observed in our sample frame across the pre, experimental, and post periods.

Our main sample frame contains surgeries that meet two criteria. First, the surgery must be one in which both the surgeon and nurse consented to participate.[14] Ultimately our participants consist of 199 surgeons from 15 surgical departments and 95 circulating nurses, in 6,251 distinct teams.[15] Second, our main sample frame includes only scheduled surgeries, which consistute 85% of all surgeries. Scheduled surgeries take place during two shifts: the morning shift (63%) and the afternoon shift (22%). The remaining surgeries (15%) are emergency surgeries; we exclude these from our main sample frame, as prespecified, since these surgeries do not receive the full intervention.[16] Finally, we note that 77% of surgeries in our sample frame constitute procedures with candidate items, which is important because the benchmarking information provided in the intervention, as well as several key outcomes (candidate costs, usage deviations), only pertain to those surgeries.

[13] In the end, the website was used only rarely.

[14] Of 261 surgeons who were contacted, 241 (92%) consented and 199 were matched in the surgery data. Of 106 nurses who were contacted, 101 (95%) consented and 95 were matched in the surgery data. See Appendix A3 for more details.

[15] Only 193 doctors were randomized. The remaining 6 were, due to a technical error, not randomized and not treated. In our main sample, we assign them to the control group. We show that our results are robust to excluding them.

[16] In Appendix **??** we additionally examine the effect of the intervention on emergency surgeries.

Figure 1 depicts the distribution of costs and item usage in the pre-period. The average surgery costs USd 440 and uses 17 disposable items, of which 12.7 units classify as candidate items. However, these statistics mask dramatic variation. Panel (a) shows the amount of all disposable items and candidate items by surgical department. The mean amount at the department level varies between less than 10 items in Eyes, Plastics and Hand specialties, to more that 50 items in vascular and heart surgeries. In virtually all departments, the lion's share of items used are candidate item targeted by our intervention. In terms of the cost of disposable equipment surgery there is a more substantial difference between the average cost of surgery which is 475.9 USD and the average cost of candidate items which is 147.17 USD. The larger gap in costs is driven by a small amount of extremely expensive items and procedures that drive the overall cost of disposable equipment.[17] Also the variations between departments is large, with average surgery costs ranging from 50 USD to over 1,500 USD in chest surgeries.

As shown in the balance tables, surgeon and nurses exhibit gender imbalance with men constituting 82% of surgeons and women 65% of nurses. In line with our description of personnel as highly skilled, both groups have substantial experience, surgeons have, on average, 18 years of experience and nurses 12 years. In line with their differential roles within the OR, surgeons and nurses differ in their degree of specialization, nurses are diverse and participate throughout our observation period in 75 distict procedures while surgeons are more specialized, with an average 17 distinct procedures. This is linked to the amount of time spent in the OR, reflected in the number of surgeries performed by each participants. Surgeries in the hospital's OR constitute the lion's share of the participating nurses' job, while most surgeons spend only a day or two a week in the OR. For both groups there is substantial variation in the amount of OR time.

Our sample frame is divided into three time periods in our analysis. The *pre-period*, October 2021 through March 2022, includes 9,078 surgeries and is the sample used for the randomization. The *experimental* period, from the end of July 2022 through the first half of May 2023, includes 14,794 surgeries and serves as the sample on which we estimate the main effects of the intervention. In Table 1 we report how these surgeries are distributed across departments, procedures, teams, and shifts. The *post-period*, second half of May 2023 through December 2023, includes 10,156 surgeries and is used to examine the persistence of our intervention.[18]

[17] The distribution of item prices is shown in Appendix Figure A4

[18] Appendix Figure A3 shows the daily frequency of the total surgical activity over the course of all three periods. This data includes non-participating personnel, emergency surgeries, and surgeries between April

2.4 Randomization

To randomize surgeons and nurses to treatment and control, we employed a matching procedure using six months of *pre-period* data, Oct 2021 to April 2022. We first stratified on the participant's field of specialization. Then, we constructed pairs of similar doctors and pairs of similar nurses using a Mahalanobis matching procedure on a vector of individual pre-period data: demographic characteristics- tenure and gender; work characteristics- number of distinct teams, number of surgeries; and measures of item usage-average deviations from the benchmark by the surgeon/nurse and average deviations from the benchmark by their teammates.[19] After constructing the pairs, each doctor and each nurse of each pair is randomly assigned to either the treatment or the control group. This results in 98 treated surgeons and 46 treated nurses.[20] Given the team structure, the personnel randomization of the personnel results in four possible treatment statuses for each surgery. Panel (d) of Table 1 shows the distribution of surgeries by the treatment status

Tables 2a and 2b show that the matching procedure succeeded in generating balance at the level of doctors and nurses, respectively. In terms of the six targeted characteristics included in the matching, all differences are small and far from statistical significance deviations from the benchmark amounts, with the exception of the number of surgeries per surgeon which is 23% lower among treated surgeons. The next five characteristics shown in Tables 2a and 2b,: native language (as a proxy for immigration and/or minority status); share of surgeries in afternoon and on-call shifts and finally our second outcome of interest-surgery costs, were not targeted but are also very well balanced across treatment and control for both surgeons and nurses.

2.5 Empirical Specification

We leverage the randomization to estimate the effects of our intervention on an outcome y_{isn} of surgery i conducted by surgeon s and nurse n, for two types of outcomes. First, at the surgery level we examine the effect of the intervention on the total cost of candidate items and the total costs of all items. At the candidate item level, we use the (standard-

and July 2022, which are not used in our analysis (during this period we recruited participants and performed the randomization).

[19] Our matching procedure followed (Lu et al., 2011) and additional details can be found in Appendix A4.

[20] After the beginning of the experiment we discovered 6 surgeon who consented to participate but because of a technical error, were not included in the randomization procedure. They are included in the analysis as part of the control groups, but the results are not sensitive to their inclusion.

ized) deviation in actual usage from the median use of the item (calculated in the pre period), for all items.[21] In order to estimate effects for surgeon treatment, nurse treatment, and the interaction, we run regressions of the form:

$$y_{isn} = \beta_0 + \beta_1 T_s^s + \beta_2 T_n^n + \beta_3 T_s^s \times T_n^n + \mathbf{X_s^s} + \mathbf{X_n^n} + \mathbf{X_i^h} + \mathbf{X_i^p} + \mathbf{X_i^t} + \epsilon_{isn} \quad (1)$$

where T_s^s is equal to 1 if the lead surgeon s in surgery i is treated and 0 otherwise. T_n^n is equal to 1 if the circulating nurse n in surgery i is treated and 0 otherwise. For each surgery i we control for the vectors of surgeon and nurse characteristics $\mathbf{X_s^s}$ and $\mathbf{X_n^n}$ that we used to balance our groups in the previous section. We also control for patient characteristics, $\mathbf{X_i^h}$, which include patient age and gender, and for all surgeries that include general anesthesia, the 1-6 ASA score. This ASA score is given by the anesthesiologist prior to surgery, representing patient's medical co-morbidities.[22] All estimates are conditional on $\mathbf{X_i^p}$ which include procedure and department fixed effects. Finally, we include a vector of time indicators, $\mathbf{X_i^t}$, which includes an indicator for the shift (morning vs. afternoon), month fixed effects, and day-of-week fixed effects. In all specifications standard errors are clustered at the surgeon and nurse level. We present q-values adjusted for multiple hypotheses using the Benjamini Hochberg procedure at the level of each table (Benjamini and Hochberg, 1995).

2.6 Treatment Implementation

To send the Whatsapp messages, we built a system that took as inputs the list of candidate items with their benchmark amounts and the schedule for upcoming shifts. Each shift's schedule, allocating surgeries (including surgeons and patients) to rooms, was set in the shift prior. Once schedules were completed, a member of the OR logistical team uploaded the schedule to our system, which automatically sent messages to corresponding treated surgeons and nurses at the beginning of the relevant shift.

Table A2 in the appendix shows the distribution of surgeries with treated personnel,

[21] Specifically, for every procedure type, we use two years of pre-period data to compute the median and standard deviation of usage for every item. To compute usage deviations, we take the number of items that were used in the surgery, subtract the median, and divide by the standard deviation. Usage deviations can be interpreted as the distance from median usage in that procedure, in units of standard deviations.

[22] The ASA Physical Status Classification System has been in use for over 60 years. The purpose of the system is to assess and communicate a patient's pre-anesthesia medical co-morbidities. For more information on ASA see https://www.asahq.org/standards-and-practice-parameters/statement-on-asa-physical-status-classification-system

by message status. In surgeries with a treated surgeon, a third (34%) of surgeons received a message prior to performing a surgery. Of these 28% matched the actual procedure performed and 6% did not.[23] In surgeries with treated nurses, 43 percent of nurses received a messages. Of these, in 17 percent of cases they received a message that did not match the actual procedure.[24] There are two main reasons why treated personnel did not receive messages at the beginning of their shift. The first is, schedule changes which are frequent both at the department level and among surgeons within the same department. Second, the messaging system that we used was paused unexpectedly several times over the course of the experiment, mainly due to changes in Facebook's terms of use for WhatsApp. This included one episode that lasted several weeks in November 2022. We leverage this pause in the experiment, which was plausibly exogenous to changes in local conditions, to characterize the effect of our intervention.

3 Results

3.1 Treatment Effects on Surgery Costs, Item Usage, and Patient Outcomes

Table 3 presents estimates of the treatment effect on surgery-level costs and item-surgery-level usage deviations from benchmark amounts. We find that treating the surgeon reduces the cost of candidate items by 9.5%, while treating the nurse has no detectable effect, whether or not the surgeon is treated too. Total surgery costs reflect a similar pattern: the effect of treating the surgeon alone is a bit smaller at 6% and not statistically significant, though it is larger and significant in surgeries where the nurse is also treated. Usage deviation allows us to examine the source of the decrease in costs—whether it is reduction of quantities (for instance, by avoiding opening items that are not eventually used) or by substituting higher cost items with lower cost ones. As with costs, we find no effect when treating nurses, either with or without treating the surgeon. For treated surgeons, we see a small increase in usage deviation of 0.03 standard deviations. However, when estimating the effect separately for items above and below the median item price (185 USD), we find that the observed increase in quantities is fully driven by an increase

[23] This is largely driven by last minute changes in patients with slightly different procedures.

[24] This is largely due to moving nurses between surgery rooms as a results of absenteeism. While specific procedures depend on the surgeon identity, nurses are more versatile, as seen in Table 2b in the number of unique procedures

in low-cost items, while there is a significant decrease in the usage deviation of more expensive items. This suggests that the treatment induced surgeons to substitute rather that reduce the overall amount of items.[25,26]

As discussed in Section 2.3, our data is characterized by a very high degree of heterogeneity, especially across different procedures, that yields substantial differences across treatment groups in the composition of surgeries. In our preferred specification we address this using a large set of fixed-effects, as presented in Equation 1. In Table 4 we present two alternative approaches to deal with this heterogeneity. First, we create a 'balanced sample', which includes only procedures (e.g. hip-replacement) that appear in both control and treatment groups, in the pre-period as well as in the experiment period, and for which there are no significant differences in costs across treatment and control in the pre-period. We see in first two columns that imposing these criteria reduce our sample size by a third, but our estimated effect on surgeons remain both large, 10 percent reduction in candidate costs, and statistically significant. In two columns on the right-hand side of the table we present results from a double machine learning procedure incorporating the set of controls in a flexible way, to better account for the heterogeneity without dropping a third of the sample. We find very similar effect sizes. Finally, as a placebo test, we estimate the treatment effects on the pre-period sample in Appendix Table A12 and find no evidence of differences across treatment groups.

Next, we test whether the intervention had any effects on patient outcomes. We run the same specification as in Equation 1 for three outcomes: the duration of the surgery, and two indicators for post-op complications: referral to post-surgery imaging and ICU referral, both within 48 hours of the surgery. The estimation results in Table 5 do not show any substantive or statistically significant effects on any of these measures. First, regarding imaging, patients require post-op imaging in 20.1% of surgeries on average (in the control group). Our point estimate for the effect of treating the surgeon is 0.2 percentage points. This coefficient is not statistically different from zero, but it is not imprecise: a 95% confidence interval rules out effects of larger than 2.9 percentage points. Given our

[25] In Appendix Table A6, we estimate our main specification on levels of costs and raw item usage, which enables us to include surgeries with zero costs and all items used in surgeries, even those without benchmark quantities. (Instead we include item fixed effects.) We see that the patterns from the main specification hold— the intervention with surgeons led to a significant reduction in the cost of candidate items, 25 USD per surgery and the substitution between expensive and cheap items is entirely driven by candidate items, with no discernible effect on the quantities of other items.

[26] In Appendix Table A3, we compute alternative p-values using a permutation testing procedure that more effectively controls Type I error in cluster randomized trials with matched pairs (Bai et al., 2024). We find very similar results.

sample size and estimated standard error, we were powered to detect effects as small as 3.9 percentage points. ICU referral is much more rare, occurring in 2.0% of surgeries in the control group. The point estimate on treating the surgeon is indistinguishable from zero, and a 95% confidence interval rules out effects larger than 0.39 percentage points. The minimum detectable effect was 0.56 percentage points. With respect to time in the operating room, we observe a slightly larger positive coefficient on treating the surgeon of 2.1 percent, but it is still statistically indistinguishable from zero, and it turns negative when both surgeon and nurse are treated. The average surgery lasted 110 minutes, so according to a 95% confidence interval, treating the doctor is very unlikely to have increased time in the operating room by more than 4.5 minutes on average. The minimum detectable effect was 5.9 percent, or 6.5 minutes.

Our intervention lasted for nearly eight months and covered 15 distinct surgical departments. This raises the concern that the overall effect is in fact concentrated in a specific time period or department. In Figure 2, we show results of an event study specification to explore the evolution of the effect over the treatment period. The effect of the doctor treatment is largest at three months; by the sixth month it is small and no longer significant. In Figure 3, we show the composition of treatment effects across departments. The effect of the surgeon treatment does not appear to be exclusively driven by one department: it is strongest in the ENT department, but has a negative sign in almost every department.

The results thus far indicate that the intervention, when targeted towards surgeons, reduced expenditure on candidate items in surgery by between 8-10%. This amounts to savings of about $13-16 per surgery, or around $20-24K monthly for an OR wing that conducts 1,500 scheduled surgeries per month.[27] These savings, driven by reduced usage of high-priced items, do not appear to have put patients at detectably higher risks of complications in surgery. In the following sub-sections, we delve into the data to provide suggestive evidence of the mechanisms behind this result. First, we make the case that the intervention worked not because item-specific information led to reduced usage of those items, but rather because it made salient the costs of disposable items in general. The dynamic effects of our intervention are consistent with this story. Second, we return to the question of the distinct roles played by surgeons and nurses, and consider how they may have interacted with the intervention.

[27] Recall that the cost of candidate items in the average surgery is around $160.

3.2 Mechanisms: The Role of Information

Our intervention includes both benchmarking information regarding the specific surgeries for the day, as well as a generic prompt against using unnecessary equipment. To what extent was the informational content important? We first exploit the fact that every message contained information only for a random set of three candidate items. In Table 6, we use our item-level data to interact treatment status with an indicator for whether the item was included in the message, and look at effects on usage deviations, both overall and for items above the median price. Strikingly, even when the item was not referenced in the message, treating the surgeon reduces usage of high priced items by around 0.1 standard deviations. We also find negative coefficients on the message interaction for both outcomes, for both surgeons and nurses, suggesting that item-specific messages may strengthen effects on usage for those items, yet standard errors are large rendering the estimates not significantly different from zero.[28]

Our results suggest that item-specific information did not drive the treatment effect. Another way to test the importance of information is to look at whether the dynamic response to treatment, both over the course of the intervention and when it stops, is consistent with a model of learning.

First, if surgeons update their beliefs in response to the information, and if this is what makes the intervention effective, then treatment effects should actually become stronger over the course of the intervention, since new messages often contain new information about items and procedures. On the other hand, if messages operate by making disposable item costs salient, we would not expect their effects to rise over the course of the intervention, and we might even expect a decline in responsiveness if personnel habituate to the messages and attend to them less over time (Embi and Leonard, 2012). We have already seen in Figure 2 that the effect of treating the surgeon definitely does not appear to increase over the course of the intervention, and if anything, it falls.

Second, if messages work by providing information, the effects should be independent of the timing of the last message received. If, on the other hand, messages make usage costs salient, then we might expect to observe patterns of "action and backsliding" that are commonly seen in the literature (e.g., Allcott and Rogers (2014)). Specifically, suppose the

[28]We also exploit the partial implementation of our intervention to check whether receiving correct information mattered at the surgery level in Appendix Table A14. We find little evidence that the receipt of correct pre-shift messages strengthened the effect of the surgeon treatment on costs.

prompt directs attention to an action, changing its marginal utility in the short run, but not in the long run as memory decays. This would cause usage to diminish, but not disappear, in the absence of messages and it would limit the persistence of the intervention effects.

To investigate lagged responses to previous messages, we first exploit plausibly exogenous failures in our messaging system, which occurred in November 2022. The failure was due to a change in Facebook's policies for WhatsApp messages, and resulted in the absence of pre-shift messages for several weeks. We replicate Table 3 for the subset of surgeries that occurred during the days when the messaging system was down, and therefore whose surgeons and nurses did not receive pre-shift messages that day. The results, shown in Table 7, suggest that lagged responses are strong in the short-run. Column 1 shows that the surgeon treatment reduces candidate costs by 12.4% even in surgeries that received no pre-shift message that day. The effect on usage deviations for items above the median price, shown in column 4, is also similar in magnitude.

Does the effect of messages exhibit similar persistence when surgeons go many days or weeks without them? First, in Figure 4 we estimate the effect of a pre-shift message on surgeries in subsequent shifts that did not receive messages. Figure 4 suggests that the response to pre-shift messages seems to wane over subsequent surgeries. These estimates make use of the fact that our intervention was imperfectly implemented, but they are limited in that very few treated surgeons went more than seven days without receiving a message. We can look at the effects of many consecutive days and surgeries without messages by examining treatment effects in the six months after the experiment ended. In Table 8 we report the results. The point estimate on the effect of treating the surgeon is 7.8% (compared to 9.5% in the experimental period), and it is no longer significantly different from zero.

All in all, the evidence suggests that the informational content of messages was not essential. The fact that sthe effects of messages wane but do not disappear over time is consistent with a model where messages direct limited attention to the costs of item usage. This attention shift, and the reduced item usage that results, persists over short periods, but not over longer ones.

3.3 Mechanisms: The Role of Autonomy

A key feature of our context, as described in Section 2.1, is the organizational structure within which surgeons and nurses operate. We illustrate the key differences in Figure 5. While both surgeons and nurses are skilled experts, surgeons have more decision-making power in the operating room, as depicted by the red arrows on the right, and face less oversight by the OR wing management, as depicted by the blue arrows on the left. For both of these reasons, surgeons can be understood to have higher autonomy than nurses.

Indeed, self-reported data from our baseline survey confirms the importance of these channels. We asked both surgeons and nurses standard questions about decision-making power in the operating room. The results, displayed in Figure 6, substantiate several facts. First, nurses are indeed considered by all to have expertise in equipment usage (subfigures (c) and (d)). Second, few disagree with the statement that physicians should make all decisions in the OR (9% of surgeons and 17% of nurses, as shown in subfigure (a)). However, when asked about relative power between surgeons and nurses, there is disagreement. Most surgeons believe it is better for nurses to agree with the surgeon, while most nurses are either neutral or hold the opposite view (subfigure (b)).

A second important difference between surgeons and nurses is management. Recall that nurses are directly managed by the OR wing administrative staff, which is responsible for the budget and cares about costs. Surgeons, in contrast, are managed by their respective departments and see themselves as simply utilizing OR services. Given this institutional structure, we might expect OR wing nurses to weigh cost considerations more highly than do surgeons at baseline. To check this, in our baseline survey, we asked four related questions: whether control of spending is a proper consideration, whether doctors should take cost-effectiveness into account, whether clinical guidelines should include cost-benefit considerations, and whether it is the responsibility of the medical provider to help control costs. Respondents had to answer between 1 (strongly disagree) and 5 (strongly agree). Figure 7 makes clear that at baseline, nurses on average place higher importance on cost considerations than do surgeons.

Both of these differences would predict that individual surgeons, who have more decision-making power and adhere less to centralized norms or rules set by management, have more varying effects on the cost of surgery than do nurses. We estimate surgeon and nurse fixed effects in a regression with candidate costs on the left-hand-side, and including the rich set of fixed effects that we use in all of our specifications on the right-hand

side. Figure 8 plots the results. There is considerably more variability in the role that doctors play in determining costs, relative to nurses, consistent both with surgeons having more decision-making power over costly inputs, and with surgeons being less influenced by a centralized management.

Can differences in autonomy explain why the intervention was effective for surgeons but not for nurses? If nurses have limited autonomy over equipment use, then ultimately they cannot be responsible for any waste due to misuse or precautionary opening. The fact that the effects of treating the surgeon were driven in part by substituting away from expensive equipment is consistent with this, since nurses are unlikely to have discretion over such substitutions. The fact that treating nurses in addition to surgeons had little added value (the interaction term is statistically indistinguishable from zero) is also consistent with this story. Furthermore, on the management side, if nurses are so well-managed that they have fully internalized the cost considerations of their supervisors, an intervention such as ours will have little leverage, even if nurses are theoretically capable of contributing to overuse.

If different levels of autonomy *across* surgeons and nurses is an important mediator of our intervention, we might also expect differences in autonomy *within* surgeons and nurses to interact with treatment effects. Exploiting the data from our baseline survey, we can look within surgeons, and within nurses, and ask whether baseline proxies for autonomy mediate responsiveness to treatment.

We first consider four variables related to decision-making power: two at the individual level, and two at the team level. The first is a "nurse power" index that aggregates (normalized) answers to the four questions shown in Figure 6, plus one more about whether nurses are members of the team.[29] We define someone as perceiving "high" nurse power if they are above the median. The second variable, nurse inclusion, is the (normalized) self-reported frequency of nurse participation in decision-making in the OR. We define someone to have "high" levels of nurse inclusion if they are above the median. The third variable is constructed by normalizing self-reported confidence for surgeons and nurses, respectively, and then taking the difference between them. The fourth variable is constructed by simply taking the difference in tenure between the surgeon and the nurse.[30] [31]

[29] We did not plot this in Figure 6 as a large majority of doctors and nurses agreed or somewhat agreed with this statement.

[30] Tenure is defined to be the number of years working in the relevant role (not necessarily at this hospital).

[31] Note that there are a few differences between the way we have defined variables here, and the way we

We also consider two variables related to baseline cost considerations. The first is an index that aggregates the four questions about the importance of cost considerations; we define someone to have "high" cost consideration if they are above the median. The second is an indicator for whether baseline usage, as measured by usage deviations, is above the median.

For each of the six attributes, we run a regression in which we interact treatment status of the surgeon/nurse with the attribute indicator for the surgeon/nurse. The outcome is log costs. The specification includes main effects for treatment status and the attribute indicator. It includes the same set of control variables as our main specification, as well as indicators for the five other surgeon/nurse attributes. The estimated interaction coefficients for the surgeon attributes are in Table 9a, and those for the nurse attributes are in Table 9b. For each specification we include only surgeries where the surgeon (nurse) answered the baseline survey.[32]

Among all the dimensions of heterogeneity, the results indicate that three are potentially important, and they all pertain to nurses. First, nurses who perceive themselves to have high power vis-a-vis the surgeon were relatively more responsive to treatment. Second, surgical teams with large tenure gaps between the surgeon and the nurse were relatively more responsive to the nurse treatment. Third, surgical teams with large confidence gaps between the surgeon and the nurse were also somewhat more responsive to the nurse treatment, though this difference is not significant at conventional levels. These general findings are robust to various alternative specifications.

All three findings suggest that in teams where the balance of power favored the surgeon, nurses were relatively *more* responsive to treatment. In other words, it is *not* the case that relatively more autonomous nurses were more responsive to treatment, but rather the opposite. It seems that our intervention was empowering for nurses who felt (relatively) inferior to the surgeon. For these nurses (and teams), the intervention reduced the cost of candidate items by only 3-5%, consistent with their influence being relatively limited, especially regarding high-priced items. However, it is possible that these nurses were, at baseline, less able to align themselves with the cost considerations of the OR wing management because they felt they had to align themselves with the interests of the surgeon, and therefore more responsive to our intervention. To the extent that the intervention

defined variables in our pre-analysis plan; see Appendix A5 for details.

[32] Almost all nurses answered the baseline survey, but many surgeons did not. Table **??** shows our main effects for the sample of surgeons and nurses who answered the baseline survey. We see that the estimated effect for surgeons on costs in this sampe is significantely smaller than in the full sample.

shifted nurse behavior (which it did not on average), it appeared to operate via an entirely different mechanism: empowerment vis-a-vis the surgeon.

We began this paper by highlighting the unique challenges of shifting the behavior of autonomous experts, and we tailored our intervention for this purpose. It may be that interventions such as ours are also uniquely suited for fully autonomous experts.

4 Conclusion

We design a non-financial intervention in attempt to reduce excessive disposable equipment use among medical experts in surgery. Because waste cannot be measured, our intervention is built to allow autonomous experts to apply it as they see fit, both in order to realize potential cost-savings that are unobserved by managers, and to avoid unintended consequences. We conduct a large field experiment, targeted not only at surgeons but also at circulating nurses, in order to understand the suitability of our intervention for experts with less decision-making power and subject to a more traditional management structure. We find that treating the surgeon reduced the cost of candidate inputs by 10% on average, translating into cost savings of about USD 26K per month. Treating the nurse was ineffective, and suggestive evidence indicates that such interventions may indeed be uniquely suited for fully autonomous experts who are not already embedded in traditional management structures.

References

Alexander, D. (2020). How Do Doctors Respond to Incentives? Unintended Consequences of Paying Doctors to Reduce Costs. *Journal of Political Economy*, 128(11):4046–4096. Publisher: The University of Chicago Press.

Allcott, H. and Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, 104(10):3003–3037.

Bai, Y., Liu, J., Shaikh, A. M., and Tabord-Meehan, M. (2024). Inference in cluster randomized trials with matched pairs. *Journal of Econometrics*, 245(1-2):105873.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Bergstresser, D. and Philippon, T. (2006). Ceo incentives and earnings management. *Journal of Financial Economics*, 80(3):511–529.

Bloom, N., Mahajan, A., McKenzie, D., and Roberts, J. (2020). Do Management Interventions Last? Evidence from India. *American Economic Journal: Applied Economics*, 12(2):198–219.

Brosig-Koch, J., Hennig-Schmidt, H., Kairies-Schwarz, N., and Wiesen, D. (2016). Using artefactual field and lab experiments to investigate how fee-for-service and capitation affect medical service provision. *Journal of Economic Behavior & Organization*, 131:17–23.

Brosig-Koch, J., Kairies-Schwarz, N., and Kokot, J. (2017). Sorting into payment schemes and medical treatment: A laboratory experiment. *Health Economics*, 26(S3):52–65. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hec.3616.

Cooper, Z., Scott Morton, F., and Shekita, N. (2020). Surprise! Out-of-Network Billing for Emergency Care in the United States. *Journal of Political Economy*, 128(9):3626–3677. Publisher: The University of Chicago Press.

Doyle, J., Abraham, S., Feeney, L., Reimer, S., and Finkelstein, A. (2019). Clinical decision support for high-cost imaging: A randomized clinical trial. *PLOS ONE*, 14(3):e0213373.

Einav, L., Finkelstein, A., and Mahoney, N. (2021). Long-Term Care Hospitals: A Case Study in Waste.

Eliason, P., Heebsh, B., League, R., McDevitt, R., and Roberts, J. (2022). The effect of bundled payments on provider behavior and patient outcomes: evidence from the dialysis industry. *Work. Pap., Duke Univ., Durham, NC*.

Eliason, P. J., League, R. J., Leder-Luis, J., McDevitt, R. C., and Roberts, J. W. (2021). Ambulance Taxis: The Impact of Regulation and Litigation on Health Care Fraud.

Emanuel, E. J., Ubel, P. A., Kessler, J. B., Meyer, G., Muller, R. W., Navathe, A. S., Patel, P., Pearl, R., Rosenthal, M. B., Sacks, L., Sen, A. P., Sherman, P., and Volpp, K. G. (2015). Using Behavioral Economics to Design Physician Incentives That Deliver High-Value Care. *Annals of Internal Medicine*, 164(2):114–119. Publisher: American College of Physicians.

Embi, P. J. and Leonard, A. C. (2012). Evaluating alert fatigue over time to EHR-based clinical trial alerts: findings from a randomized controlled study. *Journal of the American Medical Informatics Association*, 19(e1):e145–e148.

Englmaier, F., Grimm, S., Grothe, D., Schindler, D., and Schudy, S. (2024). The Effect of Incentives in Nonroutine Analytical Team Tasks. *Journal of Political Economy*, 132(8):2695–2747. Publisher: The University of Chicago Press.

Eurostat (2022). Government expenditure on health.

Gosnell, G. K., List, J. A., and Metcalfe, R. D. (2020). The Impact of Management Practices on Employee Productivity: A Field Experiment with Airline Captains. *Journal of Political Economy*, 128(4):1195–1233. Publisher: The University of Chicago Press.

Gould, D. J., Moralejo, D., Drey, N., Chudleigh, J. H., and Taljaard, M. (2017). Interventions to improve hand hygiene compliance in patient care. *Cochrane database of systematic reviews*, (9).

Hennig-Schmidt, H., Selten, R., and Wiesen, D. (2011). How payment systems affect physicians' provision behaviour—An experimental investigation. *Journal of Health Economics*, 30(4):637–646.

Howard, D. H. and McCarthy, I. (2021). Deterrence effects of antifraud and abuse enforcement in health care. *Journal of Health Economics*, 75:102405.

Kolstad, J. T. (2013). Information and Quality When Motivation is Intrinsic: Evidence from Surgeon Report Cards. *American Economic Review*, 103(7):2875–2910.

La Forgia, A. (2023). The Impact of Management on Clinical Performance: Evidence from Physician Practice Management Companies. *Management Science*, 69(8):4646–4667.

League, R. (2023). Administrative Burden and Consolidation in Health Care: Evidence from Medicare Contractor Transitions.

Leder-Luis, J. (2023). Can Whistleblowers Root Out Public Expenditure Fraud? Evidence from Medicare. *The Review of Economics and Statistics*, pages 1–49.

Lu, B., Greevy, R., Xu, X., and Beck, C. (2011). Optimal Nonbipartite Matching and Its Statistical Applications. *The American statistician*, 65(1):21–30.

Malhotra, S., Cheriff, A. D., Gossey, J. T., Cole, C. L., Kaushal, R., and Ancker, J. S. (2016). Effects of an e-Prescribing interface redesign on rates of generic drug prescribing: exploiting default options. *Journal of the American Medical Informatics Association*, 23(5):891–898.

Mohanan, M., Donato, K., Miller, G., Truskinovsky, Y., and Vera-Hernández, M. (2021). Different Strokes for Different Folks? Experimental Evidence on the Effectiveness of Input and Output Incentive Contracts for Health Care Providers with Varying Skills. *American Economic Journal: Applied Economics*, 13(4):34–69.

Navathe, A. S. and Emanuel, E. J. (2016). Physician Peer Comparisons as a Nonfinancial Strategy to Improve the Value of Care. *JAMA*, 316(17):1759.

Navathe, A. S., Sen, A. P., Rosenthal, M. B., Pearl, R. M., Ubel, P. A., Emanuel, E. J., and Volpp, K. G. (2016). New Strategies for Aligning Physicians With Health System Incentives. *The American Journal of Managed Care*, 22(9):3.

Patel, O. U., Yang, L. C., and Fazendin, J. (2023). Using Behavioral Economics to Decrease Operating Room Costs and Promote Better Surgeon Accountability. *JAMA Surgery*, 158(1):5–6.

Sacarny, A., Avilova, T., Powell, D., Williamson, I., Merrick, W., and Jacobson, M. (2023). A randomized trial of letters to encourage prescription monitoring program use and safe opioid prescribing: Study examines effectiveness of letters to clinicians that encourage prescription monitoring program use and safe opioid prescribing. *Health Affairs*, 42(1):140–149.

Sacarny, A., Safran, E., Steffel, M., Dunham, J. R., Abili, O. D., Mohajeri, L., Oh, P. T., Sim, A., Brutcher, R. E., and Spevak, C. (2022). Effect of Pharmacist Email Alerts on Concurrent Prescribing of Opioids and Benzodiazepines by Prescribers and Primary Care Managers: A Randomized Clinical Trial. *JAMA Health Forum*, 3(9):e223378.

Scur, D., Sadun, R., Van Reenen, J., Lemos, R., and Bloom, N. (2021). The World Management Survey at 18: lessons and the way forward. *Oxford Review of Economic Policy*, 37(2):231–258.

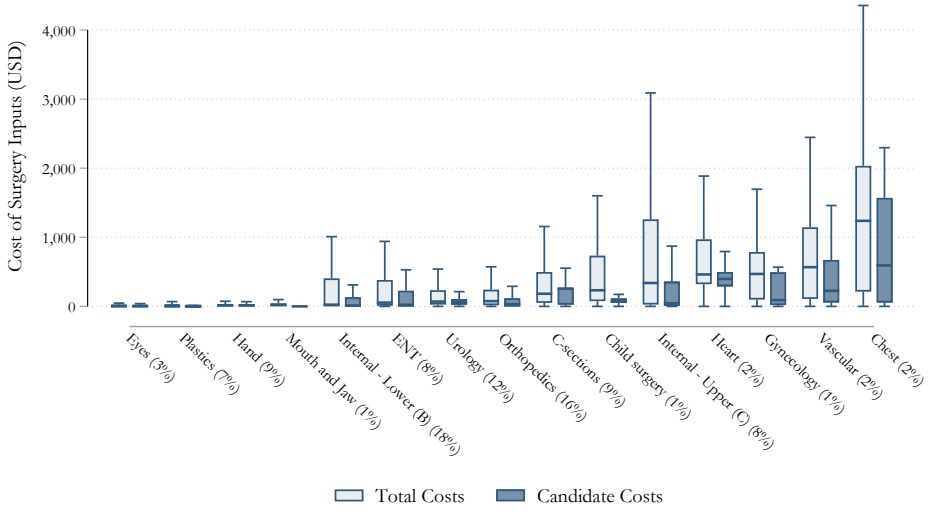
Shi, M. (2024). Monitoring for Waste: Evidence from Medicare Audits*. *The Quarterly Journal of Economics*, 139(2):993–1049.

Silver, D. (2021). Haste or Waste? Peer Pressure and Productivity in the Emergency Department. *The Review of Economic Studies*, 88(3):1385–1417.

Zureich, J. and Singh, M. (2024). Do Physicians Improve More from Positive or Negative Feedback?

Figure 1: Surgery costs and item usage by Department, Pre-Period

(a) Surgery costs



(b) Number of items

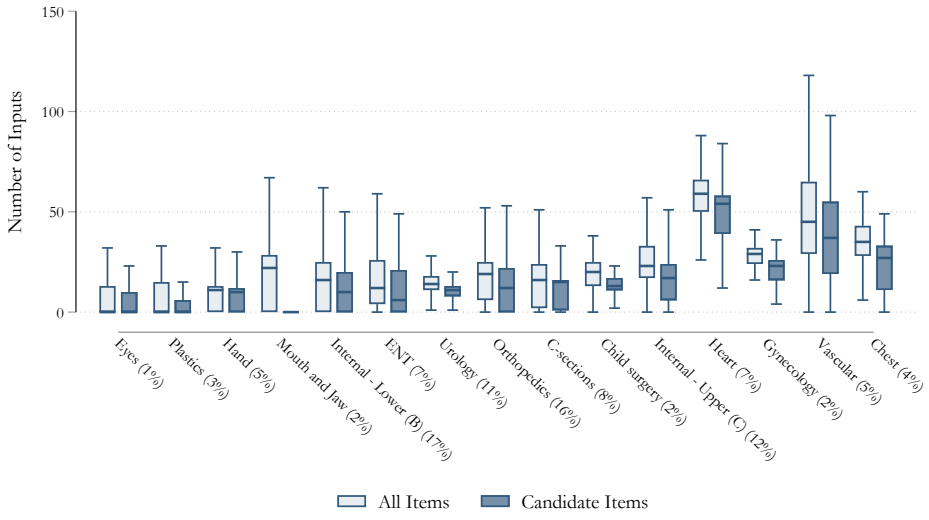
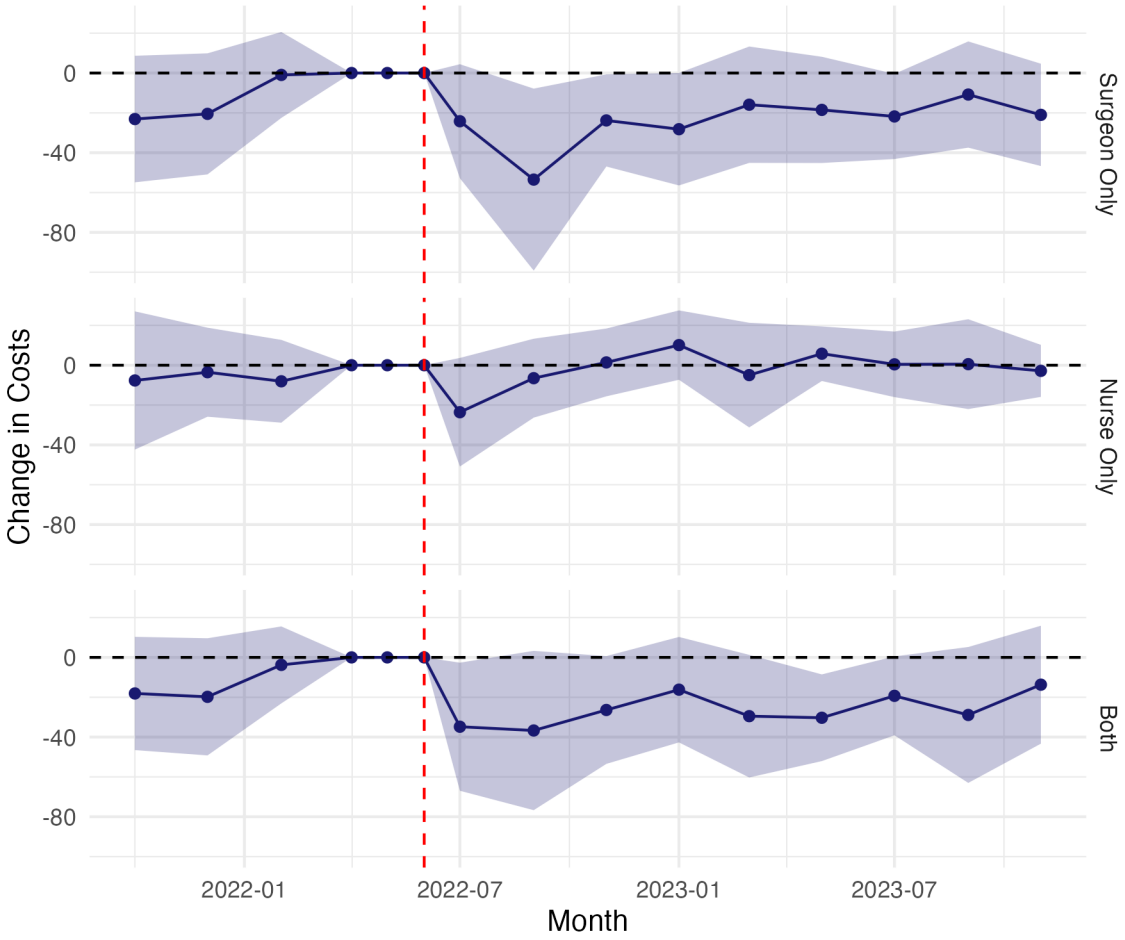


Figure 2: Event Study, Candidate Costs, bi-monthly



Estimates based on a sample size of less than 250 observations are grayed out.

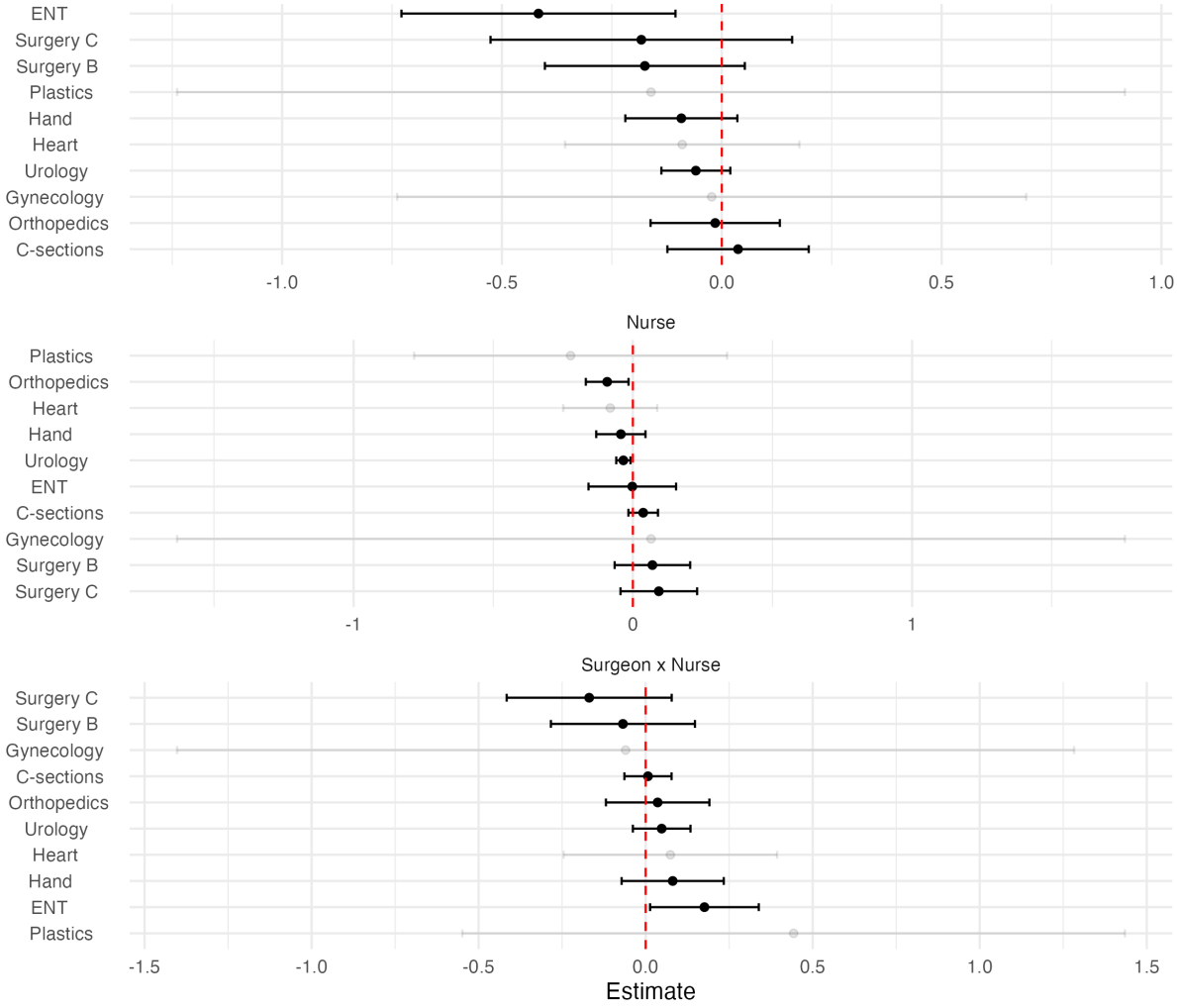


Figure 4: Dynamic Effects of Message over Surgeries

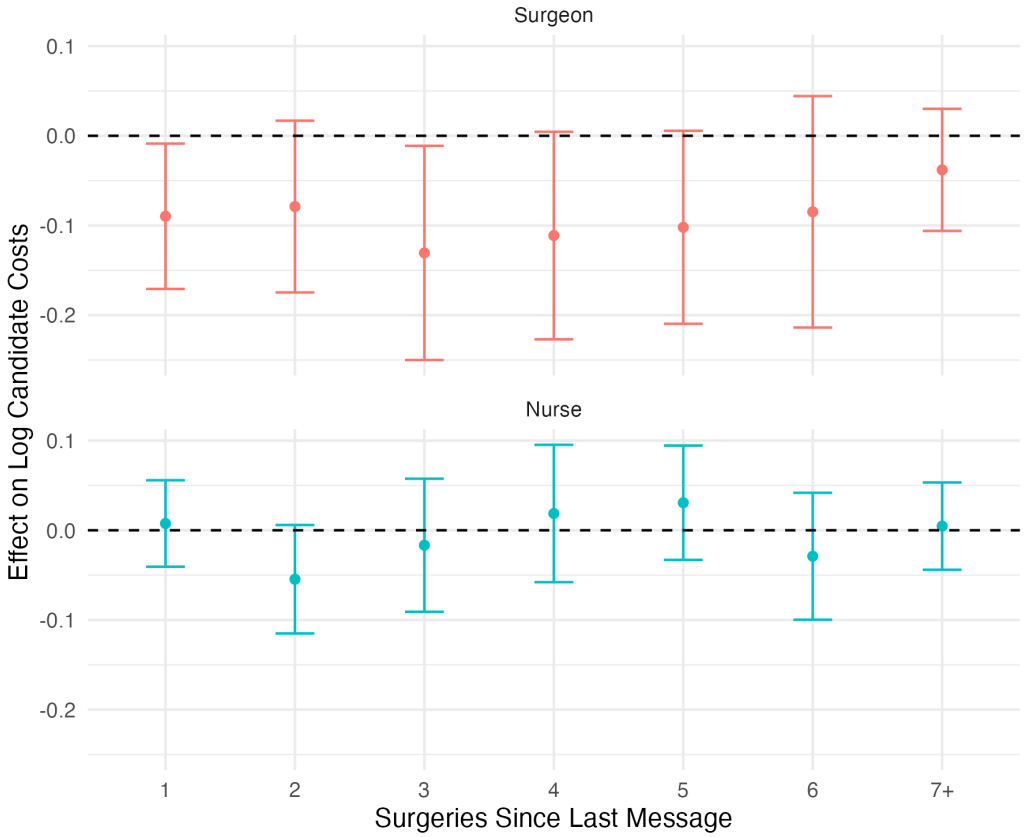


Figure 5: Organizational Structure

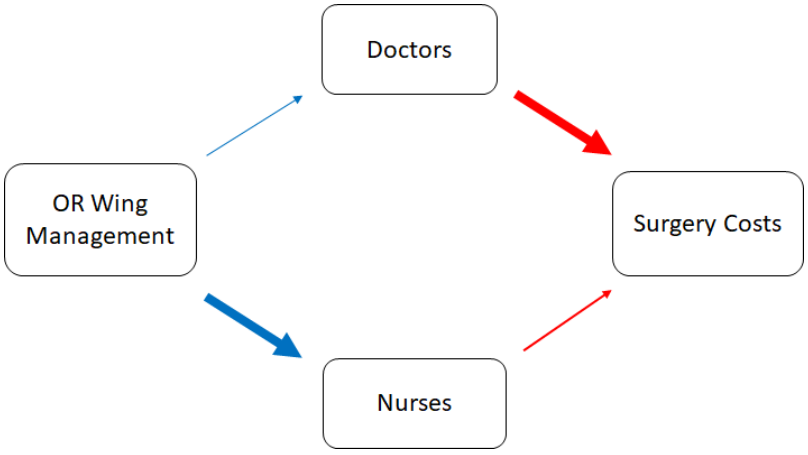
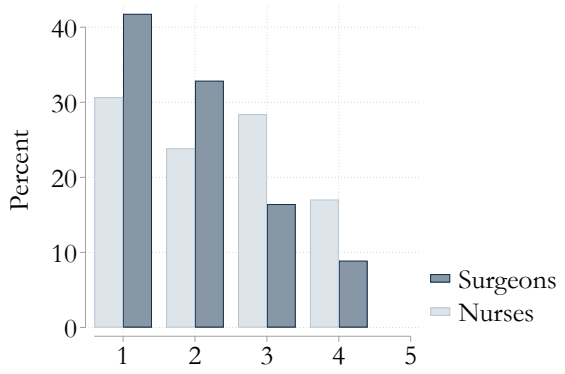
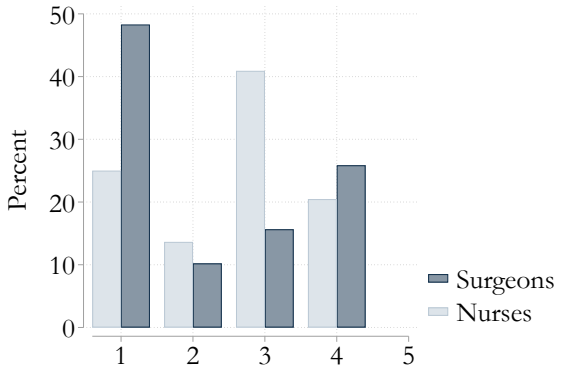


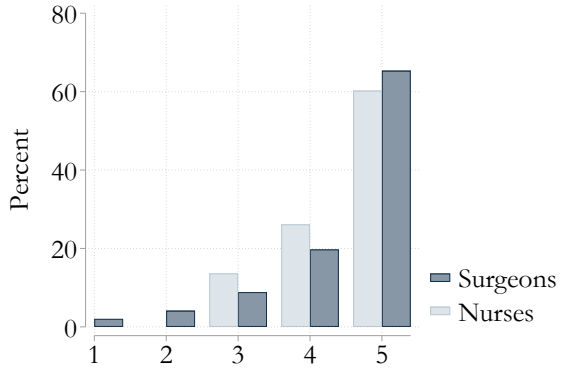
Figure 6: Baseline Perceptions of Decision-Making Power, Surgeons & Nurses



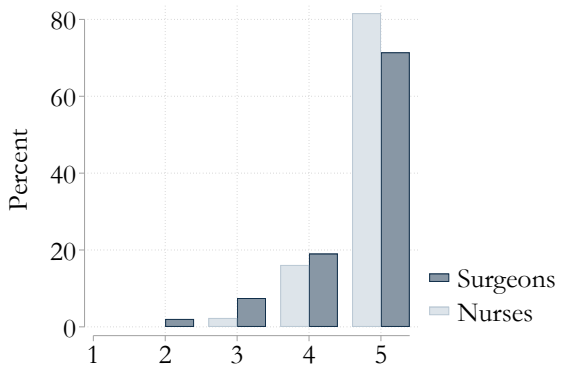
(a) Physicians should make all decisions in the OR
(1 agree, 5 disagree)



(b) It is better for nurses to agree with the surgeon
(1 agree, 5 disagree)

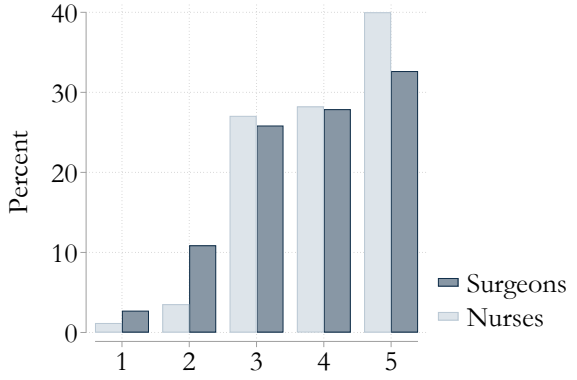


(c) Nurses are responsible for monitoring the use of equipment
(1 disagree, 5 agree)

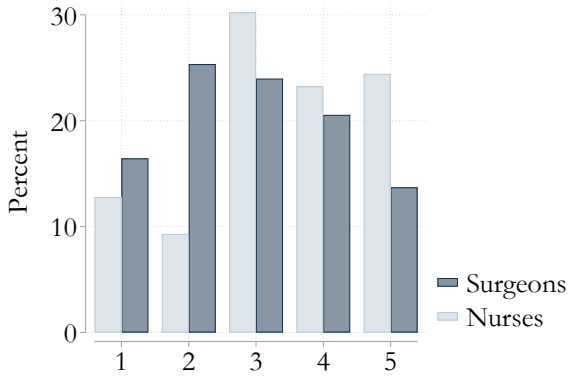


(d) Nurses have important knowledge in use of equipment in OR
(1 disagree, 5 agree)

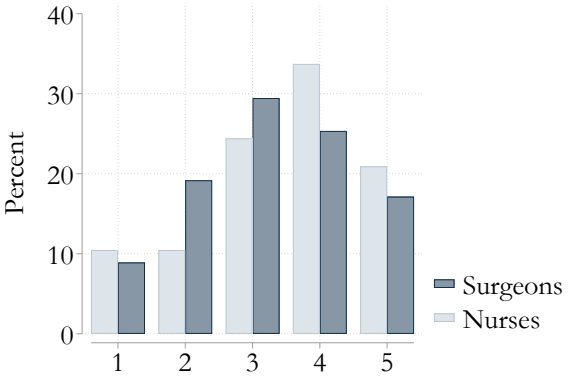
Figure 7: Baseline Cost Considerations, Surgeons & Nurses



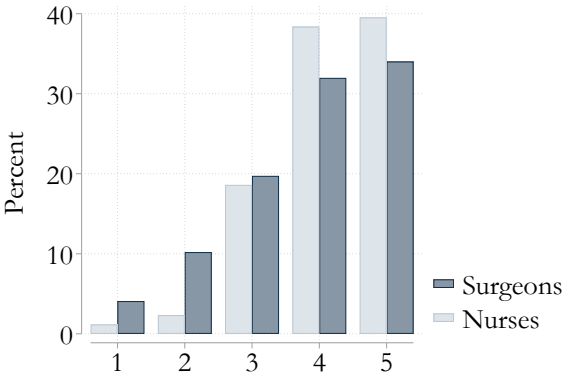
(a) Control of spending is a proper consideration
(1 disagree, 5 agree)



(b) Physicians should take cost effectiveness into account
(1 disagree, 5 agree)



(c) Clinical guidelines should include cost-benefit considerations
(1 disagree, 5 agree)



(d) It is the responsibility of physicians/nurses to help control costs
(1 disagree, 5 agree)

Figure 8: Histogram of Surgeon and Nurse Fixed Effects

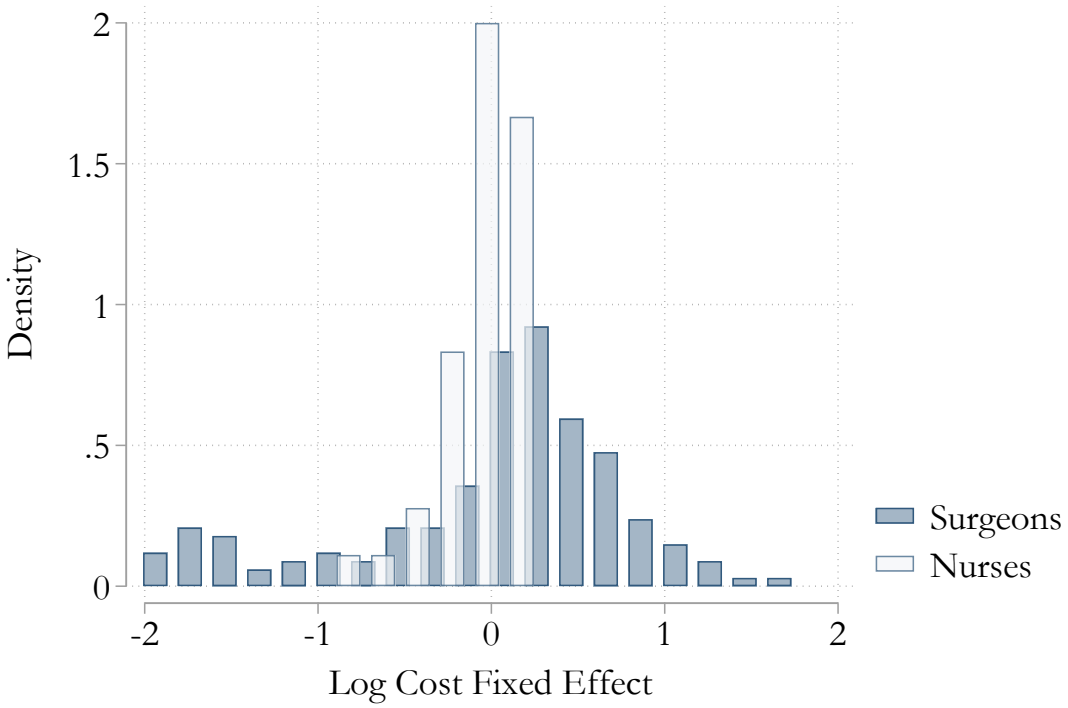


Table 1: Descriptive Statistics, Surgeries

	Total	Control	Only Surgeon	Only Nurse	Both
No. Scheduled Surgeries	12,627	3,262	3,096	3,557	2,712
No. Surgeries	14,794	3,877	3,636	4,125	3,156
Distinct Departments	15	15	14	15	14
Distinct Procedure Types	431	330	293	315	277
Fraction Surgeries w/ Candidate Items	0.77	0.77	0.78	0.78	0.75
Distinct Teams	3,913	995	1,034	927	957
Fraction Afternoon Surgeries	0.22	0.23	0.19	0.27	0.19
Fraction Emergency Surgeries	0.15	0.16	0.15	0.14	0.14

Table 2: Pre-period balance of treatment and control personnel

(a) Surgeons

	All	Control	Treated	p-value
Raw Number of Surgeries	52.58	54.97	50.13	0.53
Tenure	15.60	14.87	16.32	0.24
Male	0.78	0.78	0.79	0.94
Distinct Procedures	9.83	9.97	9.68	0.76
Average Candidate Cost	149.98	161.49	138.24	0.36
Average Total Cost	431.25	422.89	439.79	0.83
Native Lang Hebrew	0.77	0.71	0.82	0.07
Native Lang Russian	0.10	0.10	0.09	0.83
Afternoon	0.18	0.17	0.19	0.50
Emergency	0.20	0.20	0.19	0.84
Weighted No. of Surgeries	60.62	61.77	59.46	0.79
Weighted No. of Teams	23.15	22.06	24.25	0.52
Baseline Usage Deviations	-0.90	-0.90	-0.90	0.99
Baseline Usage Deviations, Team	-0.97	-0.98	-0.97	0.91
Observations	198	100	98	

(b) Nurses

	All	Control	Treated	p-value
Raw Number of Surgeries	123.82	123.52	124.11	0.97
Tenure	11.53	12.46	10.61	0.43
Male	0.29	0.33	0.26	0.50
Distinct Procedures	40.62	41.28	39.96	0.80
Average Candidate Cost	167.36	179.44	155.29	0.30
Average Total Cost	426.58	464.61	388.55	0.31
Native Lang Hebrew	0.34	0.39	0.29	0.31
Native Lang Russian	0.47	0.45	0.48	0.84
Afternoon	0.22	0.23	0.22	0.93
Emergency	0.14	0.15	0.14	0.51
Weighted No. of Surgeries	144.72	156.80	132.65	0.38
Weighted No. of Teams	51.93	57.92	45.95	0.32
Baseline Usage Deviations	-1.00	-0.96	-1.04	0.32
Baseline Usage Deviations, Team	-1.11	-1.06	-1.16	0.23
Observations	92	46	46	

Surgeries with only one participating medical professional (surgeon or nurse) are included in the balance tables but not in the analysis. Weighted number of surgeries and teams are weighted by the duration of appearance in the pre-period data

Table 3: Treatment Effects on costs and usage, Experimental Period

	Log costs Candidate	Log costs Total	Usage dev All	Usage dev >Median
Surgeon	−0.095	−0.061	0.036	−0.100
	(0.034)	(0.044)	(0.022)	(0.029)
	[0.040]	[0.416]	[0.313]	[0.009]
Nurse	−0.015	0.023	−0.001	−0.012
	(0.019)	(0.028)	(0.022)	(0.023)
	[0.725]	[0.725]	[0.968]	[0.729]
Surgeon x Nurse	0.017	−0.064	−0.016	0.008
	(0.033)	(0.038)	(0.029)	(0.037)
	[0.729]	[0.313]	[0.729]	[0.905]
Num. obs.	8,182	9,566	183,181	9,710
R^2	0.848	0.774	0.135	0.391
Mean Dep. Var.	4.312	5.122	−0.429	−0.163

Notes: Standard errors are two-way clustered at the surgeon and nurse level. Stars indicate q-values computed using the Benjamini-Hochberg step-down procedure considering all of the hypothesis within a table. * $q < 0.05$; ** $q < 0.01$; *** $q < 0.001$.

Table 4: Double machine learning results

	Balanced sample Candidate Costs	Balanced sample Total Costs	DML Candidate Costs	DML Total Costs
Surgeon	-0.100^{*}	-0.107^{*}	-0.082	-0.083
	(0.037)	(0.046)	(0.021)	(0.026)
			(< 0001)	(0.004)
Nurse	-0.005	0.032	0.002	0.010
	(0.022)	(0.037)	(0.016)	(0.022)
			(0.921)	(0.779)
Surgeon x Nurse	0.003	-0.052	-0.041	-0.049
	(0.039)	(0.046)	(0.021)	(0.028)
			(0.109)	(0.118)
Num. obs.	5457	5518	7345	7859
R^2	0.851	0.754		
Adj. R^2	0.843	0.742		
Mean Dep. Var.	4.410	5.056	4.329	5.123

Notes: Heteroskedasticity-robust standard errors, computed following the method of Chernozhukov et al. (2018), are reported in parentheses beneath coefficient estimates. Beneath each standard error, q-values are presented in parentheses, calculated using the Benjamini-Hochberg step-down procedure. The step-down procedure is applied separately to the pre- and post-period halves of the table.

Table 5: Treatment Effects on Clinical Outcomes

	Log Op Time	Imaging	ICU
Surgeon	0.021	0.002	0.000
	(0.021)	(0.014)	(0.002)
	[0.462]	[0.916]	[0.916]
Nurse	0.021	−0.005	0.005
	(0.016)	(0.014)	(0.003)
	[0.458]	[0.916]	[0.439]
Surgeon x Nurse	−0.061	−0.026	−0.003
	(0.023)	(0.016)	(0.003)
	[0.087]	[0.439]	[0.462]
Num. obs.	6,126	10,005	10,005
R^2	0.773	0.275	0.495
Mean Dep. Var.	4.698	0.201	0.020

Table 6: Treatment Effects on Usage Deviations, Experimental Period, Interaction with Message Status

	Usage Dev All Items	Usage Dev >Median Price
Surgeon	0.037	−0.098
	(0.022)	(0.028)
	[0.467]	[0.007]
Surgeon, Message for Item	−0.049	−0.071
	(0.039)	(0.109)
	[0.951]	[0.850]
Nurse	0.001	−0.010
	(0.022)	(0.025)
	[0.719]	[0.850]
Nurse, Message for Item	−0.017	−0.045
	(0.031)	(0.101)
	[0.850]	[0.850]
Surgeon x Nurse	−0.017	0.005
	(0.030)	(0.038)
	[0.850]	[0.951]
Num. obs.	183,181	9,710
R^2	0.135	0.391
Mean Dep. Var.	−0.429	−0.163

Table 7: Treatment Effects, Experimental Period, Messaging System Down

	Log Cand. Costs	Log Total Costs	Usage Dev All Items	Usage Dev $>$ Median Price
Surgeon	-0.127	-0.112	0.015	-0.127
	(0.049)	(0.075)	(0.040)	(0.051)
	$[0.093]$	$[0.335]$	$[0.870]$	$[0.093]$
Nurse	-0.010	0.011	-0.002	-0.093
	(0.030)	(0.045)	(0.030)	(0.044)
	$[0.870]$	$[0.874]$	$[0.960]$	$[0.155]$
Surgeon x Nurse	0.040	-0.030	0.019	0.092
	(0.053)	(0.076)	(0.044)	(0.055)
	$[0.870]$	$[0.870]$	$[0.870]$	$[0.298]$
Num. obs.	2,528	2,959	58,076	3,257
R^2	0.852	0.793	0.149	0.390
Mean Dep. Var.	4.440	5.266	-0.421	-0.131

Table 8: Treatment Effects on Surgery Costs, Post-Experimental Period

	Post-Period Log Cand. Costs	Post-Period Log Total Costs
Surgeon	−0.078	−0.107
	(0.050)	(0.055)
	[0.184]	[0.184]
Nurse	−0.030	0.040
	(0.024)	(0.041)
	[0.274]	[0.329]
Surgeon x Nurse	0.064	−0.098
	(0.036)	(0.060)
	[0.184]	[0.184]
Num. obs.	5,529	6,507
R^2	0.878	0.804
Mean Dep. Var.	4.320	5.188

Table 9: Heterogeneous Effects on Log Candidate Costs, Interactions with Personnel Indicators

(a) Surgeons

	Cand. Costs I = Cost Cons.	Cand. Costs I = Bsl Usage	Cand. Costs I = Nur Pwr	Cand. Costs I = Nur Incl	Cand. Costs I = Exp Gap	Cand. Costs I = Conf Gap
(I=0) x Surgeon	−0.038	−0.079	−0.049	−0.066	−0.061	−0.051
	(0.058)	(0.036)	(0.053)	(0.057)	(0.057)	(0.072)
	[0.589]	[0.309]	[0.589]	[0.589]	[0.589]	[0.589]
(I=1) x Surgeon	−0.041	−0.107	−0.052	−0.021	−0.030	−0.030
	(0.049)	(0.054)	(0.062)	(0.061)	(0.048)	(0.043)
	[0.589]	[0.309]	[0.589]	[0.729]	[0.589]	[0.589]
P-value of diff	0.957	0.647	0.971	0.567	0.522	0.706
Num. obs.	6,259	8,182	6,259	6,254	5,450	5,799
R^2	0.852	0.849	0.852	0.852	0.847	0.850
Mean Dep. Var.	4.220	4.312	4.220	4.229	4.231	4.198

(b) Nurses

	Cand. Costs I = Cost Cons.	Cand. Costs I = Bsl Usage	Cand. Costs I = Nur Pwr	Cand. Costs I = Nur Incl	Cand. Costs I = Exp Gap	Cand. Costs I = Conf Gap
(I=0) x Nurse	−0.013	−0.013	−0.050	−0.002	0.034	0.024
	(0.021)	(0.034)	(0.028)	(0.020)	(0.028)	(0.034)
	[0.800]	[0.887]	[0.291]	[0.912]	[0.620]	[0.800]
(I=1) x Nurse	−0.008	−0.016	0.043	−0.012	−0.054	−0.031
	(0.032)	(0.016)	(0.024)	(0.039)	(0.029)	(0.029)
	[0.887]	[0.620]	[0.291]	[0.887]	[0.291]	[0.620]
P-value of diff	0.887	0.920	0.003	0.822	0.013	0.172
Num. obs.	7,508	8,182	7,543	7,543	5,450	5,799
R^2	0.846	0.848	0.846	0.846	0.847	0.850
Mean Dep. Var.	4.279	4.312	4.279	4.279	4.231	4.198

Appendix

A1 Figures and Tables

Figure A1: Kits, Safe Items, Multiple Use Items



Figure A2: Pre-Shift Message, Feedback Message

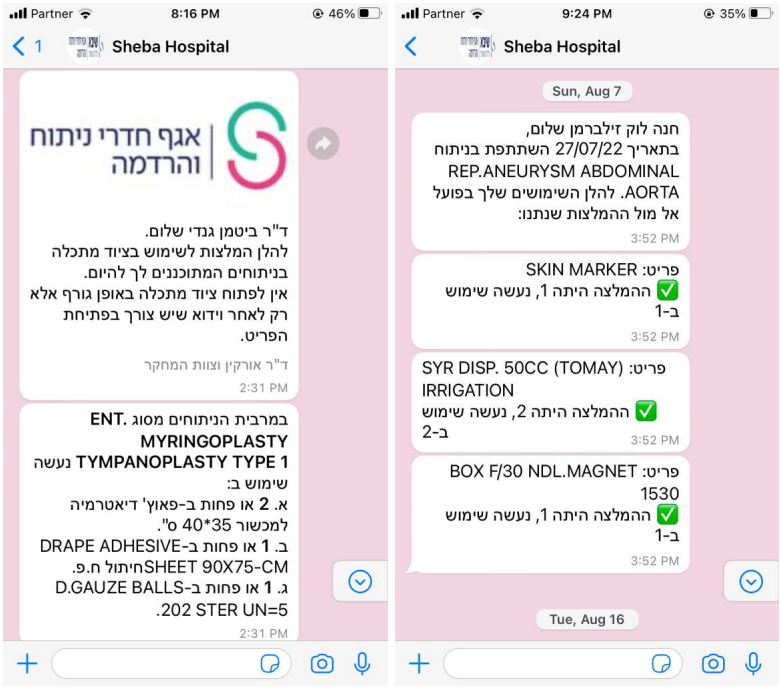


Figure A3: Surgeries by Week and Shift

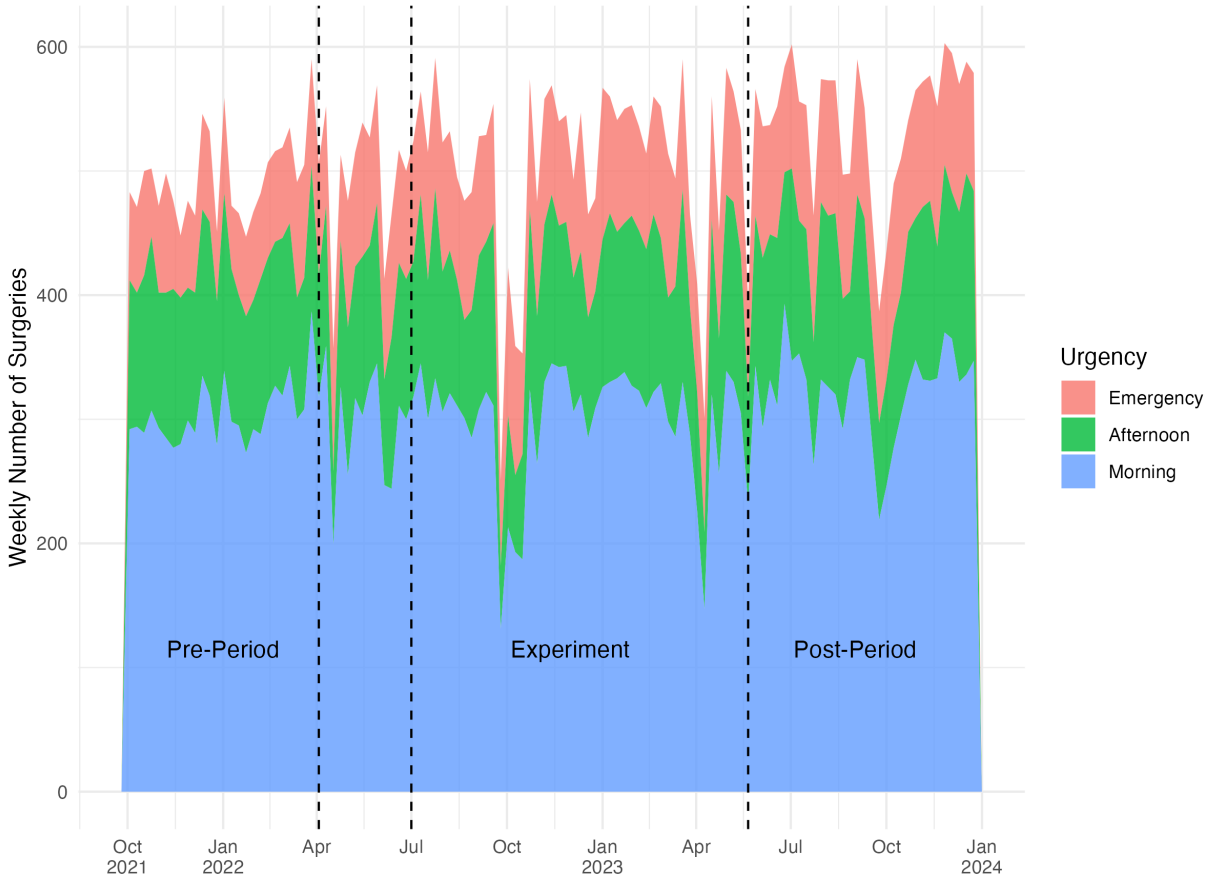
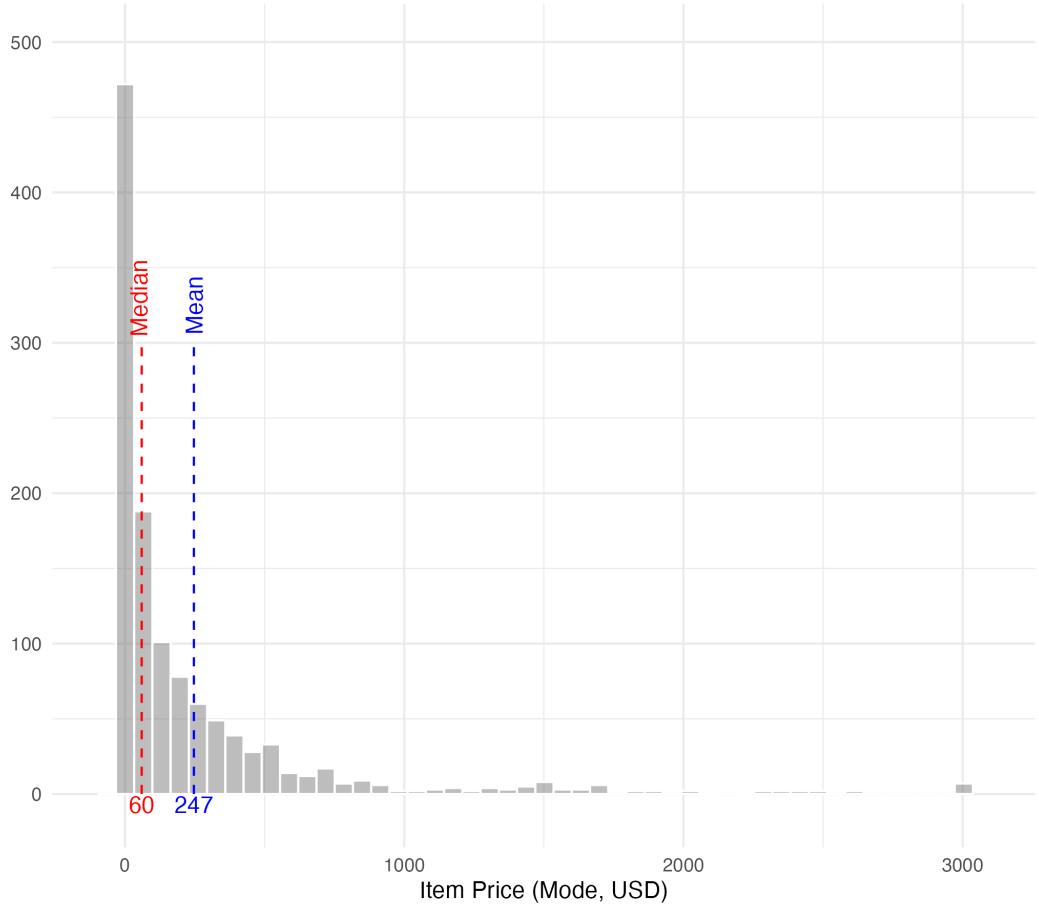


Figure A4: Candidate Items, Pre-Period



Note: Top-coded at 3000

Table A1: Pre-period balance of treatment and control, surgeries

	Control	Surgeon Only	Nurse Only	Both	p-val,joint
Total Costs	492.86	505.33	423.44	337.93	0.94
	(1,037.01)	(1,143.92)	(919.51)	(828.96)	
Candidate Costs	198.52	159.20	171.24	111.20	0.87
	(409.23)	(362.86)	(367.40)	(275.37)	
Number of items used	17.99	16.36	17.13	14.38	0.8
	(12.96)	(12.85)	(12.94)	(11.99)	
Unique items used	15.23	13.85	14.39	12.14	0.83
	(11.61)	(11.76)	(11.46)	(11.04)	
Number of candidate items used	14.56	12.45	13.56	10.30	0.75
	(14.80)	(13.37)	(14.33)	(11.39)	
Unique candidate items used	11.21	9.88	10.46	8.44	0.79
	(10.07)	(9.85)	(9.42)	(8.67)	

Table A2: Implementation of Pre-Shift Messages Among Treated Personnel

	Surgeons		Nurses	
Correct message	22.92	(1331)	20.69	(1297)
Incorrect message	5.48	(318)	14.42	(904)
No candidate items	16.44	(955)	13.62	(854)
System down	28.82	(1674)	28.60	(1793)
No message	26.34	(1530)	22.67	(1421)
Total	100.00	(5808)	100.00	(6269)

Table A3: Treatment Effects on costs and usage, permutation p-values

	Log costs Candidate	Log costs Total	Usage dev All	Usage dev >Median	Usage dev <Median
Surgeon	−0.095	−0.061	0.036	−0.100	0.046
	(0.034)	(0.044)	(0.022)	(0.029)	(0.024)
	[0.033]	[0.393]	[0.087]	[0.080]	[0.040]
Nurse	−0.015	0.023	−0.001	−0.012	0.001
	(0.019)	(0.028)	(0.022)	(0.023)	(0.023)
	[0.693]	[0.760]	[0.973]	[0.767]	[0.920]
Surgeon x Nurse	0.017	−0.064	−0.016	0.008	−0.019
	(0.033)	(0.038)	(0.029)	(0.037)	(0.031)
	[0.673]	[0.233]	[0.433]	[0.847]	[0.387]
Num. obs.	8,182	9,566	183,181	9,710	169,447
R^2	0.848	0.774	0.135	0.391	0.139
Mean Dep. Var.	4.312	5.122	−0.429	−0.163	−0.442

Table A4: Treatment Effects on costs and usage, Balanced Sample

	Log costs Candidate	Log costs Total	Usage dev All	Usage dev >Median	Usage dev <Median
Surgeon	−0.102	−0.111	0.041	−0.081	0.052
	(0.037)	(0.047)	(0.023)	(0.035)	(0.025)
	[0.107]	[0.113]	[0.223]	[0.113]	[0.136]
Nurse	−0.005	0.034	0.018	−0.006	0.022
	(0.021)	(0.035)	(0.026)	(0.027)	(0.028)
	[0.901]	[0.573]	[0.645]	[0.901]	[0.645]
Surgeon x Nurse	0.004	−0.052	−0.033	−0.010	−0.036
	(0.038)	(0.045)	(0.032)	(0.050)	(0.034)
	[0.914]	[0.573]	[0.573]	[0.901]	[0.573]
Num. obs.	5,457	5,518	130,421	7,899	118,755
R^2	0.851	0.754	0.139	0.405	0.145
Mean Dep. Var.	4.410	5.056	−0.473	−0.214	−0.488

Table A5: Treatment Effects on costs and usage, Balanced Sample with pairwise FE and clustering

	Log costs Candidate	Log costs Total	Usage dev All	Usage dev >Median	Usage dev <Median
Surgeon	−0.102	−0.104	0.074	−0.054	0.085
	(0.047)	(0.053)	(0.024)	(0.041)	(0.026)
	[0.165]	[0.205]	[0.030]	[0.414]	[0.030]
Nurse	−0.026	−0.037	0.020	−0.029	0.026
	(0.031)	(0.045)	(0.024)	(0.028)	(0.026)
	[0.480]	[0.480]	[0.480]	[0.471]	[0.471]
Surgeon x Nurse	0.020	−0.050	−0.051	−0.008	−0.057
	(0.041)	(0.041)	(0.033)	(0.060)	(0.035)
	[0.665]	[0.432]	[0.333]	[0.889]	[0.329]
Num. obs.	5,457	5,518	130,421	7,899	118,755
R^2	0.857	0.767	0.145	0.418	0.151
Mean Dep. Var.	4.410	5.056	−0.473	−0.214	−0.488

Table A6: Main specification with dependent variables in levels

	Costs Candidate	Costs Total	Item units All	Item units >Median	Item units <Median
Surgeon	−25.459	−12.323	0.002	−0.002	0.004
	(8.580)	(19.332)	(0.003)	(0.002)	(0.004)
	[0.058]	[0.788]	[0.721]	[0.721]	[0.721]
Nurse	−1.391	13.516	0.001	0.001	0.001
	(5.387)	(15.211)	(0.003)	(0.002)	(0.003)
	[0.922]	[0.721]	[0.922]	[0.823]	[0.922]
Surgeon x Nurse	−0.141	−16.553	−0.004	−0.003	−0.004
	(7.492)	(17.868)	(0.004)	(0.003)	(0.004)
	[0.985]	[0.721]	[0.721]	[0.721]	[0.721]
Num. obs.	12,600	12,600	1,148,144	206,792	901,481
R^2	0.812	0.832	0.243	0.206	0.243
Mean Dep. Var.	150.824	501.929	0.185	0.069	0.219

Table A7: Main specification excluding non-randomized surgeons

	Log costs Candidate	Log costs Total	Usage dev All	Usage dev >Median
Surgeon	−0.095	−0.061	0.036	−0.100
	(0.034)	(0.044)	(0.022)	(0.029)
	[0.040]	[0.405]	[0.301]	[0.009]
Nurse	−0.016	0.025	−0.001	−0.012
	(0.020)	(0.028)	(0.022)	(0.023)
	[0.729]	[0.729]	[0.978]	[0.729]
Surgeon x Nurse	0.019	−0.065	−0.016	0.008
	(0.034)	(0.038)	(0.029)	(0.037)
	[0.729]	[0.301]	[0.729]	[0.905]
Num. obs.	8,137	9,520	182,319	9,710
R^2	0.849	0.774	0.135	0.391
Mean Dep. Var.	4.309	5.122	−0.431	−0.163

Table A8: Main specification, ER surgeries

	Log costs Candidate	Log costs Total	Usage dev All	Usage dev > Median
Surgeon	0.135	−0.078	0.051	0.030
	(0.097)	(0.086)	(0.024)	(0.064)
	[0.836]	[0.836]	[0.465]	[0.948]
Nurse	0.005	−0.042	0.045	−0.004
	(0.075)	(0.047)	(0.046)	(0.066)
	[0.953]	[0.836]	[0.836]	[0.953]
Surgeon x Nurse	−0.072	0.007	−0.040	−0.034
	(0.104)	(0.094)	(0.057)	(0.092)
	[0.836]	[0.953]	[0.836]	[0.948]
Num. obs.	1,741	2,094	36,159	1,460
R^2	0.693	0.685	0.151	0.632
Mean Dep. Var.	3.677	4.704	−0.633	−0.797

Table A9: Common Support, ER versus Scheduled Surgeries

Surgery Type	No. Distinct Cases	No. Surgeries
Both	279	35950
Emergency Only	18	81
Scheduled Only	189	2653

This table compares candidate costs for surgeries in the "Both" category of Table A9. Means are computed first by case-emergency dummy groups, and then weighted across cases using the volume of surgeries among scheduled surgeries.

	Scheduled	Emergency
Mean Candidate Costs	170.92	107.70
No. Surg.	7,379	1,120

Table A11: Treatment Effects on Cost and Usage, Isolated System Down Days vs Consecutive

	Log Cand. Costs, Isolated	Log Total Costs, Isolated	Log Cand. Costs, Consecutive	Log Total Costs, Consecutive
Surgeon	−0.160	−0.115	−0.136	−0.113
	(0.107)	(0.148)	(0.062)	(0.079)
	[0.465]	[0.877]	[0.364]	[0.465]
Nurse	0.056	0.194	−0.001	−0.027
	(0.089)	(0.116)	(0.044)	(0.061)
	[0.911]	[0.465]	[0.983]	[0.983]
Surgeon x Nurse	0.003	−0.198	0.018	0.005
	(0.137)	(0.168)	(0.073)	(0.072)
	[0.983]	[0.576]	[0.983]	[0.983]
Num. obs.	888	1,053	1,643	1,909
R^2	0.892	0.862	0.862	0.803
Mean Dep. Var.	4.673	5.513	4.317	5.132

Table A12: Treatment Effects on costs and usage, Pre Period

	Log costs Candidate	Log costs Total	Usage dev All	Usage dev >Median	Usage dev <Median
Surgeon	−0.020	−0.026	0.022	−0.033	0.026
	(0.028)	(0.035)	(0.017)	(0.041)	(0.019)
	[0.878]	[0.878]	[0.878]	[0.878]	[0.878]
Nurse	−0.016	0.046	−0.008	−0.021	−0.006
	(0.026)	(0.033)	(0.025)	(0.039)	(0.027)
	[0.878]	[0.878]	[0.878]	[0.878]	[0.878]
Surgeon x Nurse	0.010	0.044	0.006	0.016	0.006
	(0.040)	(0.045)	(0.034)	(0.048)	(0.037)
	[0.878]	[0.878]	[0.878]	[0.878]	[0.878]
Num. obs.	5,384	6,170	123,190	6,729	114,011
R^2	0.780	0.727	0.141	0.255	0.148
Mean Dep. Var.	4.417	4.998	−0.418	−0.189	−0.430

Table A13: Treatment on costs and usage, Pre Period, Balanced Sample with pairwise FE and clustering

	Log costs Candidate	Log costs Total	Usage dev All	Usage dev >Median	Usage dev <Median
Surgeon	0.047	−0.006	0.059	−0.017	0.068
	(0.040)	(0.064)	(0.040)	(0.036)	(0.042)
	[0.964]	[0.964]	[0.964]	[0.964]	[0.964]
Nurse	−0.043	0.006	0.005	0.050	0.002
	(0.052)	(0.062)	(0.043)	(0.070)	(0.047)
	[0.964]	[0.964]	[0.964]	[0.964]	[0.964]
Surgeon x Nurse	−0.007	0.067	0.019	−0.004	0.017
	(0.072)	(0.090)	(0.048)	(0.076)	(0.053)
	[0.964]	[0.964]	[0.964]	[0.964]	[0.964]
Num. obs.	3,643	3,690	87,476	5,222	79,957
R^2	0.779	0.710	0.149	0.297	0.160
Mean Dep. Var.	4.502	4.996	−0.462	−0.303	−0.471

Table A14: Treatment Effects on Surgery Costs, Experimental Period, Interaction with Message Status

	Log Cand. Costs	Log Total Costs
Surgeon	−0.086	−0.044
	(0.032)	(0.050)
	[0.128]	[0.443]
Surgeon, Some Message	−0.042	0.096
	(0.057)	(0.075)
	[0.723]	[0.355]
Surgeon, Correct Message	0.061	−0.104
	(0.051)	(0.071)
	[0.496]	[0.358]
Nurse	−0.008	0.041
	(0.021)	(0.029)
	[0.355]	[0.128]
Nurse, Some Message	−0.045	−0.074
	(0.033)	(0.031)
	[0.432]	[0.413]
Nurse, Correct Message	0.073	0.061
	(0.033)	(0.029)
	[0.359]	[0.355]
Surgeon x Nurse	0.033	−0.044
	(0.034)	(0.042)
	[0.128]	[0.128]
Num. obs.	7,650	8,868
R^2	0.853	0.779
Mean Dep. Var.	4.341	5.142

Table A15: Treatment Effects on Surgery Costs, Experimental Period, Baseline Survey Participants

	Log Cand. Costs Original	Log Cand. Costs Surg Surveyed	Log Cand. Costs Nurse Surveyed	Log Cand. Costs Both Surveyed
Surgeon	−0.094	−0.053	−0.088	−0.034
	(0.034)	(0.046)	(0.034)	(0.048)
	[0.069]	[0.770]	[0.069]	[0.770]
Nurse	−0.014	−0.013	−0.010	−0.006
	(0.019)	(0.021)	(0.020)	(0.022)
	[0.770]	[0.770]	[0.770]	[0.775]
Surgeon x Nurse	0.016	0.032	0.032	0.016
	(0.034)	(0.039)	(0.035)	(0.043)
	[0.770]	[0.770]	[0.770]	[0.770]
Num. obs.	8,182	6,219	7,508	5,287
R^2	0.848	0.851	0.846	0.848
Mean Dep. Var.	4.312	4.226	4.279	4.237

Table A16: Heterogeneous Effects on Usage Deviations (>Median Price), Interactions with Personnel Indicators

(a) Surgeons

	Usage Dev I = Cost Cons.	Usage Dev I = Bsl Usage	Usage Dev I = Nur Pwr	Usage Dev I = Nur Incl	Usage Dev I = Exp Gap	Usage Dev I = Conf Gap
(I=0) x Surgeon	−0.021	−0.100	−0.170	−0.081	−0.114	−0.081
	(0.074)	(0.038)	(0.085)	(0.056)	(0.062)	(0.077)
	[0.777]	[0.123]	[0.210]	[0.222]	[0.210]	[0.358]
(I=1) x Surgeon	−0.104	−0.074	−0.052	−0.077	−0.073	−0.088
	(0.062)	(0.048)	(0.054)	(0.060)	(0.046)	(0.047)
	[0.217]	[0.217]	[0.374]	[0.272]	[0.217]	[0.210]
P-value of diff	0.410	0.691	0.236	0.953	0.461	0.915
Num. obs.	7,024	9,851	7,024	7,082	5,913	6,116
R^2	0.426	0.390	0.425	0.425	0.444	0.461
Mean Dep. Var.	−0.236	−0.161	−0.236	−0.234	−0.254	−0.275

(b) Nurses

	Usage Dev I = Cost Cons.	Usage Dev I = Bsl Usage	Usage Dev I = Nur Pwr	Usage Dev I = Nur Incl	Usage Dev I = Exp Gap	Usage Dev I = Conf Gap
(I=0) x Nurse	−0.056	−0.025	−0.009	−0.052	−0.034	−0.050
	(0.042)	(0.037)	(0.033)	(0.034)	(0.028)	(0.046)
	[0.489]	[0.674]	[0.868]	[0.489]	[0.489]	[0.489]
(I=1) x Nurse	0.021	−0.005	−0.040	0.041	−0.064	−0.018
	(0.023)	(0.032)	(0.033)	(0.035)	(0.056)	(0.044)
	[0.553]	[0.884]	[0.489]	[0.489]	[0.489]	[0.823]
P-value of diff	0.111	0.679	0.449	0.074	0.645	0.641
Num. obs.	8,522	9,851	8,557	8,557	5,913	6,116
R^2	0.425	0.390	0.424	0.424	0.444	0.461
Mean Dep. Var.	−0.191	−0.161	−0.191	−0.191	−0.254	−0.275

Table A17: Heterogeneous Effects on Log Candidate Costs, Interactions with Personnel Indicators, Pre-period

(a) Surgeons

	Cand. Costs I = Cost Cons.	Cand. Costs I = Bsl Usage	Cand. Costs I = Nur Pwr	Cand. Costs I = Nur Incl	Cand. Costs I = Exp Gap	Cand. Costs I = Conf Gap
(I=0) x Surgeon	0.013	−0.003	0.026	−0.005	0.028	0.035
	(0.070)	(0.041)	(0.043)	(0.051)	(0.050)	(0.067)
	[0.938]	[0.938]	[0.938]	[0.938]	[0.938]	[0.938]
(I=1) x Surgeon	0.034	−0.031	0.016	0.083	0.009	0.024
	(0.041)	(0.044)	(0.056)	(0.033)	(0.067)	(0.046)
	[0.938]	[0.938]	[0.938]	[0.180]	[0.938]	[0.938]
P-value of diff	0.797	0.679	0.879	0.192	0.735	0.882
Num. obs.	4,139	5,384	4,139	4,125	3,534	3,774
R^2	0.777	0.779	0.777	0.778	0.781	0.774
Mean Dep. Var.	4.331	4.417	4.331	4.340	4.322	4.310

(b) Nurses

	Cand. Costs I = Cost Cons.	Cand. Costs I = Bsl Usage	Cand. Costs I = Nur Pwr	Cand. Costs I = Nur Incl	Cand. Costs I = Exp Gap	Cand. Costs I = Conf Gap
(I=0) x Nurse	0.021	−0.050	−0.002	0.035	−0.008	−0.006
	(0.051)	(0.044)	(0.029)	(0.037)	(0.039)	(0.037)
	[0.980]	[0.980]	[0.980]	[0.980]	[0.980]	[0.980]
(I=1) x Nurse	−0.017	0.003	0.001	−0.052	0.011	0.017
	(0.028)	(0.031)	(0.045)	(0.036)	(0.052)	(0.042)
	[0.980]	[0.980]	[0.980]	[0.980]	[0.980]	[0.980]
P-value of diff	0.498	0.361	0.956	0.065	0.726	0.573
Num. obs.	4,913	5,384	4,925	4,925	3,534	3,774
R^2	0.780	0.779	0.780	0.780	0.781	0.774
Mean Dep. Var.	4.397	4.417	4.398	4.398	4.322	4.310

A2 Intervention Details

A2.1 Message Translations

Introductory Message

*"The recommendations for use of disposable equipment reflect the **median** amount of the item used in a specific procedure, during the last two years. Recommendations are only given for items for which there was **substantial variation** in its usage between different teams.*

For example, if the recommended amount for an item is 2, the meaning is that 2 or less units of the item were used in over half of these surgeries in the past two years. When the recommendation is 0 for a given item, the meaning is that in more than half of the surgeries on the procedure the item was not used at all.

The recommendations are not clinical recommendations, they are based on a statical analysis of actual usage patterns."

Pre-Shift Message

"Dear [name],
Please find below recommendations for the use of disposable equipment in the surgeries planned for you today. Disposable equipment is not to be opened until it is verified that the item is actually required.
Dr. Orkin and the research team"

Then, for each upcoming procedure, we provide benchmark quantities:

"In the majority of surgeries of type [procedure name] use was made of:
a. [median] or less [item name]
b. [median] or less [item name]
c. [median] or less [item name]"

Feedback Message

"Dear [name],
On [date] you participated in a [procedure name] surgery. Here are your actual items usages against the recommendation given:
a. Item: [item name]; The recommendation was [median], [actual] was used
b. Item: [item name]; The recommendation was [median], [actual] was used
c. Item: [item name]; The recommendation was [median], [actual] was used

A3 Recruitment

The recruitment of nurses to the study was carried out by the research team, who contacted the 106 OR nurses qualified as circulating nurses and invited them to participate in the experiment. Of these nurses, 101 (95%) consented and of them, 95 were matched in the surgery data. If there was no match in hospital data, it means that the nurses was not de-facto working as a circulating nurse in the 6 months prior.

For the recruitment of surgeons, the researchers together with the head of the OR wing met with the surgical department heads, presented the experiment, and obtained permission for and assistance in the recruitment of their surgeons. Only one department head refused to allow his department to participate in the experiment and they are not included in the data. Not all surgeons in each department actually perform surgeries (interns, very senior surgeons), but operationally it was more efficient to invite all staff and then use surgery data to identify the 'practicing' surgeons who are relevant for our experiment.

In our setting regular teams of surgeon-nurse are rare, and on average each couple meets 1-3 times during the pre-experiment period.

A4 Randomization Procedure

Our randomization procedure was based on the optimal non-bipartite matching procedure in **?**. For a set of N individuals, the procedure constructs the $N/2$ pairs that minimize the sum across pairs of the distance between two paired individuals. After the pairs are constructed, one of the two individuals in the pair is randomly assigned to the treatment group (and the other to the control group).

We do this procedure separately for physicians and for nurses. For physicians, we stratify by field by constructing pairs within each field.

We measure distance as the Mahalanobis distance based on the following variables:

1. The number of different teams each individual worked with during the six months prior to treatment, weighed according to the length of time the individual is observed in the six months of data.
2. The number of surgeries the individual worked with in the six months prior to treatment, weighed according to the length of time the individual is observed in the six months of data.
3. Average usage deviation in the six months prior to treatment
4. Average usage deviation by a typical individual's teammate (i.e. of the nurse expecting to work with the doctor and vice-versa).
5. Tenure: years of work as a surgeon or nurse (in any hospital)
6. Gender

A5 Deviations from Pre-Analysis Plan

In the heterogeneity analysis, we deviate from our pre-analysis plan in four ways.

1. We pre-specified looking at beliefs about prices and at attention as measured by responsiveness to a pre-experiment message. In the end we were not able to collect these variables.
2. We specified that nurse inclusion would be included in the nurse power index; here we separate it, since it seems to be capturing something different. (The nurse power questions elicit beliefs; the nurse inclusion question is about what actually happens in the OR.)
3. Instead of looking at surgeon-nurse gaps in perceptions of nurse power at the team-level, we look at the nurse power index separately for doctors and nurses. (For confidence and tenure we look at team-level differences as pre-specified.)

$$Y_i = \beta_0 + \beta_1 treat + \beta_2 costconsiderate + \beta_3 treat * costconsiderate + e$$

A6 Medical Literature

A7 Baseline Survey Representativeness

[ADD DISCUSSION OF REPRESENTATIVENESS OF BASELINE SURVEY]