The Impact of Recommendation Systems on Experts: Evidence from Physicians

Jorge Alé-Chilet, *Bar-Ilan University*
Juan Pablo Atal, *University of Pennsylvania*
Alejandra Benítez, *University of Pennsylvania*
Martín González-Cabello, *University of California, Los Angeles* *

February 20, 2022

Abstract

How do experts respond to computerized recommendation systems? We study the workers' compensation insurance program covering 40% of the Chilean population, where physicians decide whether a medical visit is work-related and therefore is fully covered by the law. We quantify effects of a computerized recommendation system on physicians' coverage decisions. Leveraging the staggered introduction of the system, we find that the alert improves physicians' decisions as measured by the extent to which coverage decision is reversed by a second-stage panel of expert reviewers. Decisions improved mostly by lowering coverage rates by 11% (4pp). A model of physician decision-making with learning and idiosyncratic preferences coupled with an analysis of physicians' heterogeneous responses show that the system alert had an effect both through information provision and by changing preferences. Our results help to understand the extent to which, and how, experts process system recommendations, which is key to assess both the desirability of providing recommendations and their optimal design.

*This paper uses confidential data from the Asociación Chilena de Seguridad (ACHS), which were handed to the authors as part of a collaborative agreement. We owe special thanks to ACHS's Data Analytics team for numerous conversations. We also thank Hannah Trachtman and conference participants in LACEA-LAMES 2021 for helpful feedback. The data used in this paper was given to the researchers fully anonymized as part of their agreement with ACHS. González-Cabello was a full time employee at ACHS until 2021.

1 Introduction

Judges, bankers, teachers, physicians, and other experts make decisions that have important consequences on other people's lives. Indeed, a petition for asylum, a loan, a grade of a large-stakes exam, or a life-threatening diagnosis can have large long-run consequences. However, those experts are potentially subject to behavioral biases or performance deficiencies like everybody else. In fact, recent papers have shown that personal biases and seemingly unrelated environmental or social events, such as warm weather, high pollution levels, or—egregiously—the outcome of a sports game, may cause large losses of future income and utility.[1]

The rise in recent years of information systems and machine learning has the potential to make experts' decisions more accurate and limit the effect of biases (Ganju et al., 2020). However, either due to the inherent limitations of prediction algorithms—which are based on historical and partial information—or due to ethical concerns, most of these systems do not make decisions on their own. Instead, the systems rely on algorithms that provide a recommendation, which the decision-maker may or may not follow (Dietvorst et al., 2018; Kawaguchi, 2021). Still, little is known about how expert decision-makers respond to data-driven recommendation systems. This knowledge is especially important in contexts where the decisions are of high stakes. Thus, understanding the extent to which, and how, experts process recommendations is key to assessing the overall desirability of providing recommendations in the first place, as well as their optimal design (Rayo and Segal, 2010; Kamenica and Gentzkow, 2011).

In this paper, we ask how experts react to computerized recommendations. Our empirical setting consists of physicians that make coverage decisions for occupational accidents insurance. We use data from one insurer that provides coverage and treatment of work-related accidents and injuries to roughly 2.6 million workers—

[1]Ingroup biases are present in judges and courts (e.g., Shayo and Zussman, 2011; Kuran and Lustig, 2012), teachers (Hanna and Linden, 2012; Lavy et al., 2018), and sport referees (Price and Wolfers, 2010; Parsons et al., 2011) Also, there is a broad literature on non-directly relevant factors that affect decision-makers. For example, judges make decisions less favorable for the defendant in warmer days (Heyes and Saberian, 2019), more favorable on the defendant's birthday, and give lengthier sentences on weeks where there was an unexpected football game loss of the main state team (Eren and Mocan, 2018). Judges, loan application reviewers, and major league baseball umpires, are influenced by past decisions (Chen et al., 2016). There is also growing evidence on people's worse performance during heavily polluted days (see Huang et al., 2020; Ebenstein et al., 2016 for evidence of investors and students, respectively). Norris (2019) show low levels of agreement between pairs of judges in refugee appeals.

around 40% of all salaried employees—in 82 clinics in Chile. Coverage is decided by primary care physicians in these clinics that, when receiving a patient, have to decide whether to classify the accident as work-related or not. A work-related accident is eligible for generous coverage: The patients get full monetary compensation of the lost workdays (compared to coverage starting on day four for regular healthcare coverage), as well as treatment and rehabilitation in the insurer's network, which is strongly preferred by workers relative to the public health-care system.

In particular, we study the effects of a data-driven recommendation system for physicians' coverage decisions, which provided doctors with coverage information based on historical data. The recommendation system was implemented as an alert in the physician's computer system—a pop-up message—that appeared just after the physician typed in a diagnosis with low historical coverage rates. Key for our econometric analysis, the alert messages were first implemented in a random subset of clinics before being rolled out to the whole clinic network.

Our main empirical analysis studies the alert's effect on physicians' behavior leveraging the alert staggered rollout. The analysis uses a difference-in-differences strategy that compares outcomes before and during the pilot in clinics that participated in the pilot with those that did not (i.e., treated and untreated clinics) with heterogeneous effect by alert diagnosis (comparing diagnoses that received the pop-up message with those that did not) while conditioning on a rich set of controls. Our results show that the alert message reduced coverage in the alert diagnoses by 11% (3.9 percentage points). This effect comes mainly from diagnoses with intermediate coverage levels, where there should be more uncertainty regarding the coverage decision. In addition, while the policy could have affected coverage of other diagnoses, we find that there were no statistically significant spillover effects. The results also show some persistent differences between treated and untreated clinics more than a month after the end of the pilot. Moreover, there is no evidence for statistically different trends between treated and untreated clinics, which lends credibility to our empirical design.

A common hurdle in evaluating empirically the performance of experts in their decisions is the absence of an observable measure of the true outcome. A distinctive feature of our setting is that, after the physician makes the coverage decision, the case is reviewed by a small committee of experienced physicians and assistants in each clinic that may potentially overturn the physician's decision. We find that

the alerts reduced the overturning decisions of the committee by 63% (2 percentage points). This finding shows that the alerts were effective in achieving the management's objectives and improving the quality of decisions [2].

In the last part of the paper, we leverage physicians' heterogeneity to shed light on the mechanisms by which the alert affects coverage decisions. Following the persuasion literature (DellaVigna and Gentzkow, 2010) we aim to understand whether the alert persuaded physicians about coverage decisions through an information provision channel or through preferences. We motivate this exercise with a model of physician decision-making with learning and physician idiosyncratic preferences. Our empirical test of the model consists of studying the heterogeneous effects of the alert by physician experience and by physician coverage preferences on both physician coverage and on correction probability. We find evidence for both mechanisms at play.

Contribution

This paper contributes to a growing strand of the literature that examines the effects of information technology (IT) adoption in the context of health care. Among others, Athey and Stern (2002) evaluates the effect of IT adoption in emergency calls systems on patients outcomes; Agha (2014) and McCullough et al. (2016) estimate the impact of health information technology adoption on patients' outcomes; Epstein and Ketcham (2014) studies the impact of IT on physicians' prescribing decisions; and Bundorf et al. (2019) analyzes the effect of AI recommendation system on health insurance choice (Medicare Part D).[3] We focus on automatized information provision to physicians. Close to our paper, McNamara (2021) finds that the implementation of an alert that reminds physicians that a patient is likely to need colorectal cancer testing is effective to improve screening tests. Our contribution relative to McNamara's is that we are able to examine how physicians responded relative to the impartial committee's assessment, and that we examine the mechanisms by which physicians responded to the alert. Other recent work, such as Ribers

[2]While the committee knew that the pilot was implemented, the committee did not know which case was receiving the alert and which was not. The fact that we find decreases in correction only for alert diagnoses indicates that it is driven by an improvement in physician's coverage decision and not by a change in overall correction for the committee

[3]Also, a broad set of papers study the effects of adoption of electronic medical records (Parente and McCullough, 2009; McCullough et al., 2010; Miller and Tucker, 2011) or clinical decision support systems (Ganju et al., 2020) on patients' outcomes, hospital quality, and discrimination.

and Ullrich (2019, 2020); Hastings et al. (2020), shows with simulations that machine learning recommendations could improve upon physicians' decisions. In contrast, our paper studies the implementation of such recommendation systems in a real setting, which is something those papers suggest doing.[4]

Some recent studies analyze the potential role of data-driven methods in assisting decision-making. Dietvorst et al. (2015, 2018) show in the lab that the take up of recommendations is higher if decision-makers can modify the algorithms themselves. Kawaguchi (2021) analyzes the implementation of a new recommendation system and examines the conditions under which retail workers are more likely to adopt algorithmic recommendations. We study the adoption of an algorithmic recommendation on physicians in a setting which has large stakes for patients.

This paper also contributes to a vast work of empirical papers that analyze the effects of communication. This work, related to theoretical literature that is concerned about the design of information provision (e.g., Kamenica and Gentzkow 2011), seeks to quantify the effects of communication on behavior and to understand the mechanisms through which it operates (Gentzkow and Shapiro, 2006; DellaVigna and Kaplan, 2007; Martin and Yurukoglu, 2017; Dubois and Tunçel, 2021). Our quasi-experimental setting allows us to examine the mechanism through which the message acted. We distinguish between information- and preference-based models (DellaVigna and Gentzkow, 2010). This approach is related to Ackerberg (2001), which performs a similar exercise in advertising.

2 Institutions

Since 1968 Chilean law requires all companies to provide insurance for work-related accidents. Such workers' compensation insurance is provided either by a public insurer or by non-profit private organizations (mutualidades) the largest of which is ACHS ("Asociación Chilena de Seguridad"), our partner organization. Private organizations cover close to 80 percent of the workforce of the country, and ACHS

[4]We study the implementation of a binary algorithmic recommendation as opposed to giving the decision-maker the probability calculated by the algorithm. Binary or discrete recommendations are quite common and are used in media platforms (Hallinan and Striphas, 2016), in recent implementations of food labels (Araya et al., 2019; Alé-Chilet and Moshary, 2020), and in quality certification of restaurants (Dai et al., 2018), and health insurance plans (McCarthy and Darden, 2017).

covers half of them, i.e., more than 2.6 million workers.[5]

Workers' compensation insurance includes full coverage of medical expenses of work-related accidents and illnesses, including physical and occupational therapy, and the full salary for the period when the worker is on medical leave. Work-related accidents covered by worker's compensation insurance include accidents that relate to their job duties or employment, occupational illnesses that are directly caused by work-related activities, and accidents that happen while commuting from and to the workplace. Workers can also get a pension for permanent disability due to work-related accidents, depending on how severe the disability is. Also, workers tend to prefer treatment in the *mutualidades* than treatment in regular public insurer clinics.[6]

Coverage eligibility depends mainly on the primary care physician, who assesses whether the medical event fits in the categories defined by the law. In particular, eligibility depends on how the accident happened, the employee's previous health conditions, and the type of work that the employee does.

A feature of ACHS is that after physicians make their coverage decisions a committee reviews the cases and may decide to overturn the physicians' decision. This committee is composed of the clinic's head physician, the administrative manager, and an assistant. The committee meets daily and reviews the decisions of the previous day. The committee receives a list of the cases ordered by an algorithm, and it reviews the cases until the time allocated for reviewing ends. The committee's decisions are binding so that the coverage decision process ends with the committee's assessment.

2.1 The alert experiment

The coverage decision is not a medical decision. Therefore, especially new and young primary care physicians are inexperienced in coverage eligibility.[7] To aid

[5]The public insurer is the "Instituto de Seguridad Laboral" (ISL). The other mutualidades are: Mutual de Seguridad C.Ch.C. and Instituto de Seguridad del Trabajo (IST). The default option for new companies and for independent workers is the public insurer. In practice, most small and medium companies never change insurer. For more details and statistics about the system, see the Appendix and `https://www.suseso.cl/608/w3-propertyvalue-59606.html` [October 30th, 2021]. Also, less than 1% of companies self insure.

[6]67% of workers were very satisfied with the service provided by the mutualidades ("Satisfación Mutualidades Segmento Trabajadores.´´ GFK Adimark, January 30, 2016. `https://www.suseso.cl/607/articles-496723_archivo_01.pdf` Accessed on-line February 15, 2022.)

[7]ACHS has a physician churn rate of over 50% per year because many physicians leave ACHS to do their medical specializations.

physicians in their coverage decisions, ACHS introduced a message in the form of a pop-up in the physicians' computer interface that appeared at the moment when the physician entered selected diagnoses codes. The message appeared whenever the diagnosis entered had a large historical probability of not being work-related. That probability was calculated using a boosting gradient algorithm that used data of coverage from 2018 to June 2019. The alert message read: "According to historical decisions of physicians in ACHS, the chosen diagnosis has a high probability of being ineligible for coverage." The Appendix shows a computer screenshot of the alert.

Importantly, ACHS tested the message performance with a staggered introduction. In a first stage, that began on November 8, 2019, ACHS implemented the alert in 25 of the 82 clinics of the network and was then rolled out to all of them on February 17, 2020. The choice of the clinics that participated in the first rollout (that we refer to herein as treated clinics) was random.[8]

3 Data

We use administrative data on the universe of admissions of ACHS health care providers from January 2018 to September 2020. Most of our analyses use data until February 17, 2020, the date on which the experiment ended.[9] During our sample period, the monthly number of admissions laid between 11,000 and 15,000. Each admission record includes ACHS's internal diagnosis code (2,473 unique codes groups) from which we obtain International Classification of Diseases (ICD-10) codes. The records also include the physician and the review committee coverage decision, patient characteristics (gender and age), clinic identifiers, and physician identifiers and demographics (gender, age, tenure at ACHS). After dropping 134 physicians who worked in both treatment and control clinics during the experiment period, our sample includes 690 physicians with around 45 visits per month.[10]

[8] According to ACHS the randomization "tried to balance observed characteristics such as the number of physicians and patients in each of the groups." We see this as stratification. We do not find divergent trends between treated and untreated clinics in either alert and non-alert diagnoses as we explain in Section 5

[9] Due to the COVID-19 pandemic we mostly do not use data for the post-experiment period. The pandemic had a large effect on the scope and composition of work-related accidents. The first COVID case in Chile was reported on March 3, 2020.

[10] We drop those physicians to reduce the possibility of contamination between treated and not treated clinics. More precisely, we keep physicians that saw at least 95% of their patients in either

Table 1 presents a description of cases admitted in ACHS by patients and physicians characteristics for treated and control clinics, and for and non-alert diagnoses. Panel (A) presents cases by patient characteristics and Panel (B) shows cases by physician characteristics. The table shows that cases admitted in our sample have similar patient characteristics in terms of gender and age between treated and control clinics, and between alert and non-alert diagnoses. On the other hand, we see that overall physicians' gender, age, and tenure, are similar between control and treated clinics, and alert and non-alert diagnosis, but present slight differences in terms of nationality between control and treated clinics. In our empirical strategy, we use physicians' fix effect, correcting by each physician general level of coverage.

Table 1: Summary statistics

	All	**Clinic**		**Diagnosis**		**Control**	**Treated**
		Control	**Treated**	**No alert**	**Alert**	**+ Alert**	**+ Alert**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
				A. Patient characteristics			
Female (%)	38.0	37.1	39.4	35.8	40.1	39.3	41.3
Age	40.4	40.3	40.4*	39.7	40.9*	40.7	41.3*
Observations	66,496	37,509	28,987	31,624	34,872	20,748	14,124
				B. Physician characteristics			
Female (%)	28.0	29.3	25.5*	28.1	27.8	30.3	24.8*
Age	33.9	33.9	33.9	33.9	33.9	33.8	34.1
Tenure	17.3	19.0	17.0*	17.4	17.4	18.8	16.8*
Observations	261	146	115	143	118	66	52

Note: Summary statistics based on our main sample. The table shows patient and physician demographics across clinics and alert groups. Control clinics correspond to clinics where the alert was implemented. Alert diagnoses correspond to diagnoses for which the alert was shown within the treated clinics. Statistical differences between even and odd columns in odd columns using starting in column (3). *** p<0.01, **p<0.05, * p<0.1.

As mentioned above, the recommendation alert was implemented at treated clinics and for a selected group of diagnoses with low historical coverage rates. Figure 1 shows the distribution of physician coverage decisions in our sample for alert and non-alert diagnoses for average coverage rates lower than 99%. Panel (a) presents

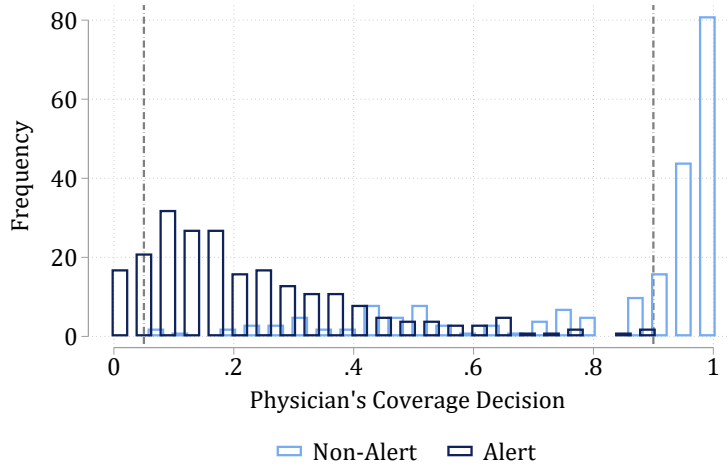
treated or nontreated clinics.

the diagnoses distribution and Panel (b) the visit distribution. The figure shows that indeed alert diagnoses tend to have low coverage rates during the pre-pilot period. Correspondingly, non-alert diagnoses tend to have larger physician coverage rates.

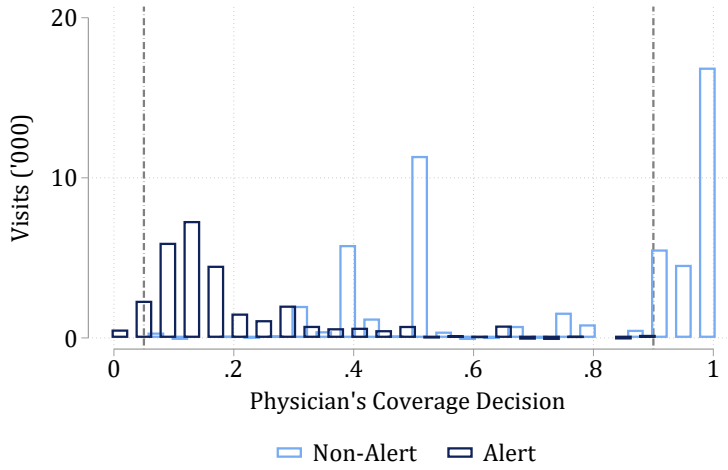
We focus our analysis on diagnoses for which there is an overlap between alert and non-alert diagnoses as measured by their historical coverage probability because the alert is more likely to affect the decision on these diagnoses. Thus, we keep diagnoses whose historical coverage probabilities are within 5 and 90%. Figure 1 shows such overlap: The vertical line shows the overlap region.

Table 2 compares the coverage decision according to patient and physician characteristics before the alert experiment. Around 33% of admitted cases in the baseline database were given coverage by ACHS physicians. Coverage was similar over patients' gender and slightly higher for younger patients, but the difference is not significant at 10%. On the other hand, near 2.5% of the cases that received coverage by the physician in the first instance were reversed by the committee secondary revision. This correction was similar for different patients' characteristics. In addition, the table shows that younger physicians give a 1–2 percentage point lower coverage on average than older ones, although the clinic committee's correction is larger than the former group only for physicians between 30 and 50 years old. Probably related with the age, physicians with lower tenure shows 2 percentage point lower coverage, while the correction from the committee for both groups is similar. Finally, there is no (or very small) difference in coverage and correction over physicians' gender, but a large difference in coverage by nationality. National physicians give on average around 35% coverage in the sample, while foreigners, 29%. Nevertheless, we don't see a relevant difference in the level of correction that both groups receive.

Figure 1: Physician Coverage Rates by Alert Inclusion



(a) Diagnoses



(b) Visits

Note: The figure shows the distribution of pre-treatment average coverage rates. An observation in panel (*a*) is a diagnosis code and in panel (*b*) is a visit. For visual purposes, the figure includes diagnosis codes with more than 20 visits, and it does not include 129,634 visits (out of 216,171) for which the average coverage decision is equal or higher than 0.99. The vertical lines show the overlap region between alert and non-alert diagnoses.

Table 2: Summary statistics - Coverage

	Patient characteristics			Physician characteristics		
	Coverage	**Correction**	**Obs.**	**Coverage**	**Correction**	**Obs.**
All	33.3	2.5	66.496	33.3	2.5	66.496
Women	33.0	2.8	25.286	33.4	2.7	22.851
Men	33.4*	2.4*	41.210	33.2	2.4*	43.645
<30 years	33.1	2.5	15.952	31.5	2.4	66.496
30 to 50 years	33.5	2.5	33.014	34.2*	2.8*	12.418
≥50 years	32.9	2.5	17.530	35.8*	2.1	33.397
Tenure < 2 years	-	-	-	32.2	2.5	945
Tenure ≥2 years	-	-	-	34.2 *	2.5*	32.097
National	-	-	-	35.2	2.6	34.399
Foreign	-	-	-	29.1 *	2.3*	45.488

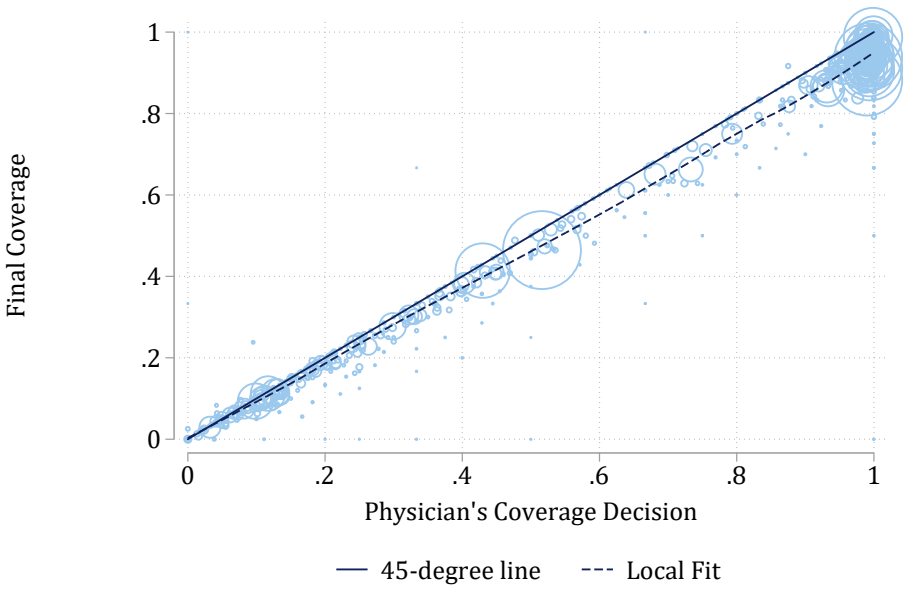
Note: Summary statistics based on our main sample. Each observation is one visit. Statistical difference with respect to base group (women, <30 years, Tenure < 2 years, and National, respectively). *** p<0.01, **p<0.05, * p<0.1.

Finally, Figure 2 presents the final coverage decision—after the committee correction—plotted against the average physician coverage rate for each diagnosis code. Each circle in the plot is a diagnosis code, where its size is the number of the admitted cases for each diagnosis. The figure includes a local quadratic fit (dashed line). For comparison, the figure also shows a 45-degree (solid) line. The figure shows that the clinics' revision committee generally revised the initial decision given by physicians from coverage to no coverage. Also, decision reversals happen with a larger probability for diagnoses codes with larger physician coverage rates.

4 Empirical Strategy

Our goal is to understand how the alert message changed physicians' coverage decisions. In particular, the alert may have had a direct effect of the alert on coverage decisions, but it may also have had potential spillovers to non-alert diagnoses if physicians infer that the absence of an alert is a message in itself. To that aim, we exploit the rollout of the alert experiment over time and over clinics and examine the alert effect on both alert and non-alert diagnoses. Hence, our main empirical specification corresponds to a difference-in-differences (DiD) model with heterogeneous effects on alert and non-alert diagnoses. The model allows for such heterogeneous

Figure 2: Correction of Physicians Coverage Decision



Note: The figure plots the final coverage rate against the initial coverage rate, a 45-degree line, and a quadratic local fit. An observation in the plot is a diagnosis code weighted by the number of visits. The figure includes only pre-pilot visits.

effects by including interaction terms between treated group and treatment period, with indicators of alert and non-alert diagnoses.

Formally, let Y_{ijct} denote the coverage decision of physician j for visit i at clinic c in week t. The main specification is as follows:

$$\begin{aligned} Y_{ijct} = &\beta\, Treated_c \times Post_t \times Alert_i + \gamma\, Treated_c \times Post_t + \\ &\nu X_i + \delta_t + \mu_i + f_j + \epsilon_{ijct}, \end{aligned} \tag{1}$$

where *Treated* is a dummy variable that indicates whether the clinic was part of the experiment, *Alert* indicates whether the visit was part of the alert experiment, *Post* is a dummy that indicates the experiment period, X is a vector of patient-visit characteristics (a quadratic polynomial of age, gender, and employer's 1-digit Standard Industrial Classification, SIC), μ are diagnosis-group fixed effects, δ are month$\times$year and day-of-the-week fixed effects, f are physician fixed effects, and ϵ is a random shock.[11] The parameters of interest are β, the effect of the message on alert diagnoses, and γ, the effect on non-alert diagnoses.

[11]The physician fixed effects absorb almost all the variation across clinics.

The main identification assumption is the parallel trends assumption of difference-in-differences models. The assumption is that there are no differential trends between treated and non-treated clinics for both alert and non-alert diagnoses in the absence of the experiment.We check for divergent trends before the intervention between treated and non-treated clinics for alert and non-alert diagnoses using the following dynamic model:

$$Y_{ijct} = \sum_{\tau=1}^{T} \beta_\tau Treated_c \times \mathbf{1}_{\{\tau=t\}} \times Alert_i + \sum_{\tau=1}^{T} \gamma_\tau Treated_c \times \mathbf{1}_{\{\tau=t\}} + \nu X_i + \delta_t + \mu_i + f_j + \epsilon_{ijct} \tag{2}$$

We do not reject parallel trends if the β_τ and the γ_τ coefficients are zero for the pre-intervention periods.

5 Results

5.1 Main Specification

This subsection presents the results of the treatment effects on physicians' decisions. All specifications in the paper cluster standard errors at the clinic level because clinics were the unit of treatment (Abadie et al., 2017). Also, all specifications include month×year and day-of-the-week fixed effects, patient characteristics (a quadratic polynomial on patient age, patient's gender), physician (or clinic) and 1-digit SIC fixed effects.

Table 3 presents the results of the main model. Columns (1) use the sample of all cases, and Columns (2)–(5) use the main sample that excludes the diagnoses with extreme average coverage rates. Also, the columns present different specifications with different fixed effects and patient demographics. The results indicate a significant effect of -2.8% of the alert message (-3.9% + 1.1%): Physicians reduced the coverage they provided by 8 percentage points, from 35% to 32.2% as a result of the alert. The effect mainly comes from the main sample where there is overlap between alert and non-alert diagnoses. Also, we do not see any spillover effect on non-alert diagnoses.

Table 3: Treatment Effects on Physician Coverage Decision

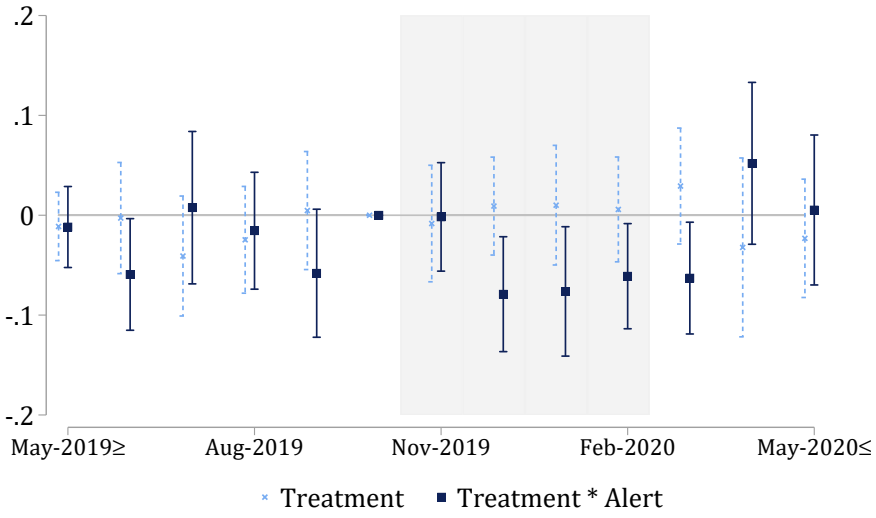
	(1)	(2)	(3)	(4)	(5)
Treated * Post * Alert	-0.020	-0.037**	-0.038**	-0.039**	-0.039**
	(0.013)	(0.018)	(0.018)	(0.018)	(0.018)
Treated * Post	-0.000	0.006	0.007	0.012	0.011
	(0.005)	(0.019)	(0.019)	(0.019)	(0.019)
Patient					X
Time F.E.	Post	Post	Month	Month	Month
Clinic F.E.	X	X	X		
Physician F.E.				X	X
N	263043	64296	64296	64294	64294
No. Clusters	82	80	80	80	80
Mean Dep. Var.	0.82	0.35	0.35	0.35	0.35
R-Squared	0.63	0.22	0.22	0.24	0.24
Sample	All	Main	Main	Main	Main

Note: The table shows the results of the DiD model. All specifications include diagnosis group, month×year, and day-of-the-week fixed effects; and patient characteristics (quadratic polynomial on patient's age, and patient's gender and employer's 1-digit SIC fixed effect). Standard errors clustered at the clinic level * $p<0.10$, ** $p<0.05$, *** $p<0.01$

5.2 Dynamic Effects

Figure 3 shows the point estimates and the 95% confidence intervals of the triple-difference coefficients of Equation (2) for every month. The figure shows no significant effects before the beginning of the alert experiment, which indicates no differences between alert and non-alert diagnoses before the messaging experiment, and thus no differential pre-trends. Also, there is a significant negative effect starting in December 2019, two weeks after the beginning of the experiment. Finally, the figure shows some persistent differential effects on treated clinics (between alert and non-alert diagnoses) one month after the experiment ended, which disappeared when the COVID pandemic began to unfold.

Figure 3: Dynamic Effects on Physician Coverage



Note: The figure shows the point estimates and the 95% confidence intervals of the estimation of the dynamic DiD models (Equation 2) for the main sample. The regressions include diagnosis group, month×year, and day-of-the-week fixed effects; and patient characteristics (quadratic polynomial on patient's age, and patient's gender, and employer's 1-digit SIC fixed effect). Standard errors are clustered at the clinic level. The grey area indicates the experiment period.

5.3 Placebo Checks

This subsection presents various placebo checks of our empirical strategy. Table 4 shows the results. Columns (1) and (2) change the definition of the pilot period to July 8, 2019, to November 8, 2019 (Column 1); and April 8, 2019, to July 17, 2019 (Column 2), while dropping the pilot period from the sample. Column (3) defines treated clinics randomly so that the share of visits in both real and fake treated clinics is similar (46%). Specifications in Columns (4) and (5) choose the alert diagnoses are chosen randomly as those that were treated. The model in Column (4) uses roughly the same number of unique alert diagnoses as in the intervention (289), and Column (5) uses roughly the same number of visits with the alert diagnoses (16 thousand). None of the estimated coefficients is significant in the placebo models.

Table 4: Placebo Checks

	Timing		Clinic	Diagnoses	
	(1)	(2)	(3)	(4)	(5)
Treated * Post * Alert	0.008	-0.015	0.006	0.000	-0.006
	(0.018)	(0.014)	(0.014)	(0.007)	(0.005)
Treated * Post	-0.002	-0.003	0.004	0.000	0.002
	(0.004)	(0.003)	(0.004)	(0.006)	(0.005)
N	220406	220406	263043	263043	263043
No. Clusters	82	82	82	82	82
R-Squared	0.63	0.63	0.63	0.61	0.61

Note: The table shows the results of the triple differences specification of placebo checks. All specifications include diagnosis group, time, patient characteristics, and physician fixed effects; and month×year and day-of-the-week fixed effects. Columns (1) and (2) do not include the intervention period. Standard errors clustered at the clinic level * p<0.10, ** p<0.05, *** p<0.01

6 The Effect of the Alert System on Decision Quality

The previous section showed that the alert system changed physicians' behavior towards granting less coverage when the alert appeared. However, to evaluate the desirability of the alert system we are ultimately interested in whether the alert system *improved* the quality of doctor's decisions.

Typically, evaluating whether the decision of a decision-maker was either right or wrong is challenging, as the "truth" is unknown to the researcher. We can get a more approximate measure of the true outcome exploiting the fact that the physician's decision is reviewed by an experts' committee review. Using the data on the outcome of the committee review, we can evaluate whether the alert system improved the quality of the physician's decision by assessing the effect of the alert on whether the physician's decision was reversed upon review.

Formally, we define $\tilde{Y}_{ijct}$ to be an indicator variable for coverage by the expert panel. We estimate a similar specification to that of equation (1), where the dependent variable is $D_{ijct} = (\tilde{Y}_{ijct} \neq Y_{ijct})$ where Y_{ijct} is an indicator variable for physician coverage. That is, D_{it} indicates cases when the physician's decision was overturned in the secondary inspection. More specifically, we assess whether the alert message changed the accuracy of the coverage decision, as measured by the

instances in which the expert panel corrects a coverage decision and deems the case as not eligible for coverage.[12]

Table 5 reports the results. We find that the probability of correction of physicians' coverage for alert diagnoses decreases as a consequence of the implementation of the alert message. The results are large: the alert reduces secondary inspection corrections by 30 percentage points (1.9 - 0.9% from a baseline of 3%).

Table 5: Correction Probability

	(1)	(2)	(3)	(4)	(5)
Treated * Post * Alert	-0.016***	-0.019***	-0.019***	-0.019***	-0.019***
	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
Treated * Post	0.005	0.008	0.008	0.009	0.009
	(0.004)	(0.006)	(0.006)	(0.006)	(0.006)
Patient					X
Time F.E.	Post	Post	Month	Month	Month
Clinic F.E.	X	X	X		
Physician F.E.				X	X
N	263043	64296	64296	64294	64294
No. Clusters	82	80	80	80	80
Mean Dep. Var.	0.04	0.03	0.03	0.03	0.03
R-Squared	0.02	0.02	0.02	0.02	0.03
Sample	All	Contested	Contested	Contested	Contested

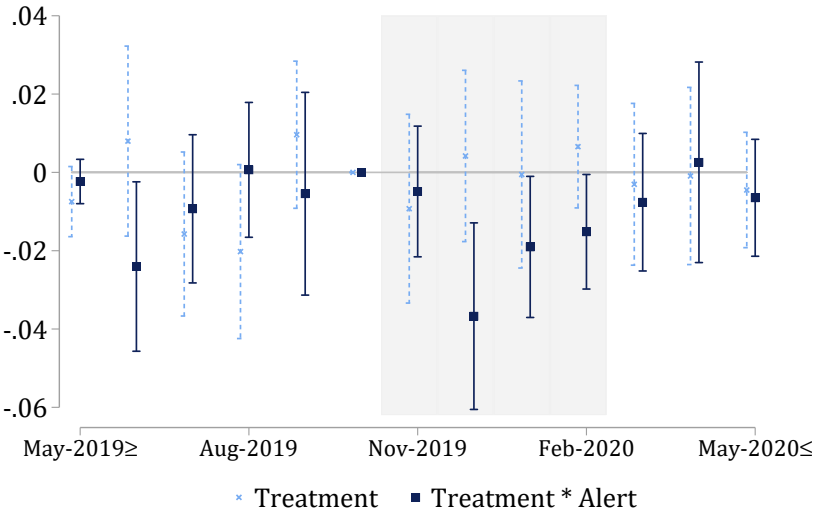
Note: The table shows the results of the diff-in-diffs and the triple differences specification. All specifications include diagnosis group, month×year, and day-of-the-week fixed effects; and patient characteristics (quadratic polynomial on patient's age, and patient's gender and employer's 1-digit SIC fixed effect). Standard errors clustered at the clinic level * $p<0.10$, ** $p<0.05$, *** $p<0.01$

As we did in Section 5, we also show a dynamic specification for the correction probability in a secondary inspection. We present the results in Figure 4. We observe no difference in the coverage of alert diagnoses between treatment and control clinics before the intervention. After the alert implementation, we see a decrease in the correction probability of the coverage of alert diagnoses in treated clinics compared

[12]In our sample the committee overturns mainly positive coverage decisions. Cases where non-coverage decisions by physicians are overturned by the expert panel are extremely rare (only 0.3% of visits).

to control ones.

Figure 4: Correction Probability



Note: The figure shows the point estimates and the 95% confidence intervals of the estimation of the dynamic DiD models (Equation (2)) for the full sample when the dependent variable is the probability of a occurrence of a secondary correction. The regressions include diagnosis group, month×year, and day-of-the-week fixed effects; and patient characteristics (quadratic polynomial on patient's age, and patient's gender and employer's 1-digit SIC fixed effect). Standard errors are clustered at the clinic level. The grey area indicates the intervention period.

7 Why Did the Alert Affect Physicians' Decisions?

In this section, we discuss the potential mechanisms behind the change in physician behavior induced by the alert system. Following DellaVigna and Gentzkow (2010) we distinguish between an information-based and a preference-based model to explain doctors' responses to the message. Under an information-based model, physicians derive utility from deciding correctly so that the physician only changes behavior if the message is informative, that is, if the message changes the physician's assessment on whether the case is indeed eligible for coverage or not.[13] Alternatively, under a preference-based model, compliance with the message enters

[13]Physicians' utility function captures altruism towards the insurer and potential future rewards stemming from compensation, reputation, career benefits, etc.

directly into the physician's utility function, as in Stigler and Becker (1977) or Becker and Murphy (1993). In this case, a message may change a physician's behavior even if it does not convey useful information.

7.1 Theoretical model

Formally, we consider instances where the physician does not know with certainty whether the case merits coverage or not. The true state of case i is a latent discrete unobserved variable Y_{it} that is drawn from a Bernoulli distribution, with the probability of being a work-related accident equal to p_i.

The utility that physician j gets from granting coverage to case i in period t (relative to not covering the case, which we normalize to 0), depends on the physician's expectation about p_i; the physician's own preferences, which we capture through a physician-specific parameter that increases the utility of coverage by λ_j; and an idiosyncratic shock ν_{ijt}. The parameter λ_j accounts for everything that makes a given physician more prone to provide coverage than others, such as altruism towards the patient (for example, the doctor wants to provide coverage to make patients happy), or other factors that could include different perceptions regarding monetary and non-monetary costs of deviating from the truth.

We therefore write the utility of providing coverage (relative to not coverage) as

$$U_{ijt} = E[p_i|\Omega_{ijt(s)}] + \lambda_j(s) + \nu_{ijt}.$$

The term $\Omega_{ijt}(s)$ is the information that physician i has about case i in period t. That information depends on the number of cases the physician has seen in the insurer's clinics and on any feedback that she had received on decisions made in the past. The term $\Omega_{ijt}(s)$ also includes the signal coming from the alert message s.

Consider a physician who has seen N_{it} cases prior to case i. Among those, the physician has provided coverage to C_i and rejected R_i cases. Covered cases are split into "successful" C_i^s and "unsuccessful" C_i^u cases, such that $C_i = C_i^u + C_i^s$. A successful covered case is one such that the doctor did not receive any negative feedback regarding the coverage decision, whereas an unsuccessful covered case is such that the doctor did receive such negative feedback.[14] Similarly, we split rejected cases as

[14]We interpret this feedback in a general way, that is, this feedback may include feedback from the expert panel but also any other means in which the physician may become aware of a mistaken

$R_i = R_i^u + R_i^s$. Let the prior distribution of p be beta distributed with parameters a_0 and b_0, so that the mean of expected coverage (without information) is equal to $\mu_0 = \frac{a_0}{a_0+b_0}$. Then, the mean of the posterior distribution after seeing N_j cases is

$$\mu_i = \frac{a_0 + C_i^s + R_i^u}{a_0 + b_0 + N_i}$$

Note that as N_i increases, the prior becomes less important and μ_i approximates $\frac{C_i^s + R_u^i}{N_i}$, which is the average rate of cases that merit coverage.

Consider now the information given in the intervention. The alert message provides information about the coverage of historical cases, which includes a large sample of cases. Let the total number of cases across physicians be N_T, of which, as before, C_T^s have been successfully covered and R_T^u have been unsuccessfully rejected. We denote C_{-i}^s the number of successfully covered cases by physicians other than i such that $C_T^s = C_i^s + C_{-i}^s$ and R_{-i}^u the number of unsuccessfully rejected cases by physicians other than i such that $R_T^u = R_i^u + R_{-i}^u$. The mean of the posterior distribution after incorporating the information of past decisions from ACHS physicians is equal to

$$\mu_j' = \frac{a_0 + C_i^s + R_i^u + C_{-i}^s + R_{-i}^u}{a_0 + b_0 + N_i + N_{-i}}.$$

Then, the effect of the message on the mean of the posterior distribution is the difference of the posterior with and without the message, $\mu_i' - \mu_i$, which is equal to

$$\mu_i' - \mu_i = \frac{C_{-i}^s + R_{-i}^u}{a_0 + b_0 + N_i + N_{-i}} - \frac{N_{-i}}{a_0 + b_0 + N_i + N_{-i}} \times \frac{a_0 + C_i^s + R_i^u}{a_0 + b_0 + N_i}.$$

Consider first a case where $N_{-i} \gg N_i$, which can be the case of a diagnose that is very common and a physician that has low experience in the insurer. In this case, the alert, that provides information about a very large number of cases relative to the physician's cases, will add relevant information to the physician. In that scenario, the first part of the second term in the expression above is close to 1, and then we have

$$\mu_i' - \mu_i|_{N_i' \gg N_i} \simeq \mu_i - \left(\frac{R_{-i}^s + C_{-i}^u}{a_0 + b_0 + N_i + N_{-i}} - \frac{a_0 + C_i^s + R_i^u}{a_0 + b_0 + N_i} \right).$$

decision.

Intuitively, when $N_i' \gg N_i$, the alert corrects the mean of the posterior distribution by the difference between the physicians' believe pre-alert and the historical level of coverage of ACHS' physicians. That is, if the physician has low level of information, the alert will correct her beliefs of level of coverage towards the historical coverage of the insurer.

Conversely, consider the case when $N_{-i} \simeq 0$, that is, the alert provides information about a very small number of cases. This could be the case of a diagnose that is not common, so the physician know that the historical coverage consider a few number of cases and doesn't give her accurate information about the real level of coverage. Then, $\mu_i' - \mu_i|_{N_i' \simeq 0} \simeq 0$.

The expression above formalizes that, overall, the signal should have a lower effect on the posteriors when the doctor is more experienced.

In addition to the effect through new information, the model allows for the signal to change the physician's preferences for coverage even if it does not change the posterior of p. This could happen if, for instance, physicians have a preference for complying with the message or if the message changes their assessment of the potential cost of deviating from the true state. Formally, we could think of the signal as a cost of providing more coverage than a given threshold. Thus, the average coverage from physician j is

$$\bar{y}_j = E[E[p_i|\Omega_{ijt(s)}] + v_{ijt}] + \lambda_j(s).$$

If cases are randomly allocated to doctors with different preferences, we can write

$$\bar{y}_j = \mu + \lambda_j(s), \tag{3}$$

with $\mu \equiv E[E[p_i|\Omega_{ijt(s)}] + v_{ijt}]$ is the average rate of cases that merit coverage in the population.

Consider a message that generates a cost to physicians with $\bar{y}_j > K$. From equation (3) it is easy to see that physicians with a higher λ, those who tended to give more coverage, will react more to the message. The message in the case will lower the level of λ, so that the fix term in the utility for giving coverage will be lower after receiving the message than before.

How to distinguish empirically whether the message affected behavior through information or through preferences? On the one hand, an informative message

should affect more physicians that are less informed ex-ante (e.g., less experienced physicians because they have a weaker prior.) On the other hand, a message under a preference-based model should change physicians' behavior even when the message conveys no information or even when physicians have a strong prior. In particular, under a preference-based model, it is likely that the message affects more those doctors with an ex-ante higher taste for coverage.

7.2 Empirical tests

We test the predictions of the model by looking at the heterogeneous effects of the message along two dimensions: ex-ante information and ex-ante coverage preferences. To this aim, first, we proxy for the degree of physicians' ex-ante information with their tenure with the insurer (as measured by years since the start of her contract, log visits, and log visits in each diagnosis). Note that the message depends on historical coverage rates for each diagnosis, so the extent to which the information was already available to the physicians should naturally vary with the physician's experience.

Second, we proxy for physicians' coverage preferences with physician-specific coverage propensities. We estimate those propensities as the physician fixed effects in the estimation of the following regression:

$$Y_{ijct} = \nu X_i + \delta_t + \mu_{d(i)} + f_j + \epsilon_{ijct}, \tag{4}$$

where, as in Section 4, Y_{ijct} is the coverage decision of physician j for visit i at clinic c in week t, X_i is a vector of patient/visit characteristics (a quadratic polynomial of age, gender, and employer's 1-digit Standard Industrial Classification, SIC), $\mu_{d(i)}$ are diagnosis-group fixed effects, δ_t are month$\times$year and day-of-the-week fixed effects, f_i are physician fixed effects, and ϵ_{ijct} is a random shock. The estimating sample includes the pre-intervention period and physicians with more than 30 visits only.[15]

Table 6 presents evidence in favor of the preference-based mechanism and against the information-based mechanism. The table shows the effect of the intervention on alert diagnoses interacted with demeaned physicians' characteristics. Columns (1)–(3) present the test of the information-based mechanism. The columns show the

[15]Figure A2 in the Appendix shows the estimated fixed effects and a binned scatter plot of those fixed effects against physician tenure.

Table 6: Heterogeneous effects by Physician Characteristics on Physician Coverage

	(1)	(2)	(3)	(4)	(5)
Treated * Post	0.013	0.013	0.010	0.012	0.013
	(0.019)	(0.019)	(0.019)	(0.018)	(0.018)
Treated * Post * Alert	-0.040**	-0.019	-0.043**	-0.030	-0.029
	(0.018)	(0.022)	(0.019)	(0.019)	(0.019)
Treated * Post * Alert * Tenure	-0.016***				-0.012**
	(0.005)				(0.006)
Treated * Post * Alert * ln(Visits)		-0.033**			
		(0.016)			
Treated * Post * Alert * ln(Visits-Diag)			-0.009		
			(0.010)		
Treated * Post * Alert * Coverage				-0.042***	-0.039***
				(0.009)	(0.008)
N	62315	62315	62315	62245	62245
No. Clusters	79	79	79		
R-Squared	0.24	0.24	0.24	0.24	0.24

Note: The table shows the results main specification where the DiD coefficients are interacted with physician characteristics. The variable *Coverage* represents physician coverage propensity calculated as described in the text. The sample includes only physicians with more than 30 visits in the pre-intervention period. All specifications include diagnosis group, time, patient characteristics, and physician fixed effects; and month$\times$year and day-of-the-week fixed effects. Bootstrapped standard errors in Columns (4) and (5). Standard errors clustered at the clinic level * $p<0.10$, ** $p<0.05$, *** $p<0.01$

results of interaction effects with physicians' tenure at ACHS (measured by years working for ACHS, number of visits, and number of visits by diagnosis, respectively). Contrary to what would be implied by the information-based mechanism, we find that doctors who reacted the most are those with higher experience.

Column (4) shows the test of the preference-based mechanism. The column presents results that include the interaction of the intervention on alert diagnoses with the physician coverage propensity. The interaction with physician propensity is significant, which indicates that physicians who had a higher reduction in coverage with the alert message were those who had a larger propensity to give coverage in the baseline period. Column (5) shows that these results are robust to including both an interaction with tenure.

8 Conclusions

This paper examines the effect of an alert recommendation system implemented in a large network of clinics that delivers health care for work-related accidents. The alert provides information to primary-care physicians, who have to decide whether the case is work-related and coverage-eligible or not. We find a reduction in the level of coverage given by primary care physicians for the alert diagnoses. The effect comes mainly from diagnoses with intermediate historical coverage levels, where there should be more uncertainty regarding the coverage decision. Importantly, we also find that the alert lowered correction rates of physicians' decisions by the clinic's management. We interpret our results in light of the persuasion literature and find evidence for a preference/persuasive effect in play.

One issue we were not able to address in the paper is the optimal alert design (as in, e.g., Vatter, 2021). For example, the alert was binary in the sense it provided physicians with coverage/no coverage information. In principle, a better design could tell physicians the exact probability of coverage by their peers, although this may produce information overload. We leave these issues for future work.

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey Wooldridge (2017) "When should you adjust standard errors for clustering?"Technical report, National Bureau of Economic Research.

Ackerberg, Daniel A (2001) "Empirically distinguishing informative and prestige effects of advertising," *RAND Journal of Economics*, pp. 316–333.

Agha, Leila (2014) "The effects of health information technology on the costs and quality of medical care," *Journal of Health Economics*, Vol. 34, pp. 19–30.

Alé-Chilet, Jorge and Sarah Moshary (2020) "Beyond Consumer Switching: Supply Responses to Food Packaging and Advertising Regulations," *Available at SSRN 3678744*.

Araya, Sebastian, Andrés Elberg, Carlos Noton, and Daniel Schwartz (2019) "Identifying food labeling effects on consumer behavior," *Available at SSRN 3195500*.

Athey, Susan and Scott Stern (2002) "The impact of information technology on emer-

gency health care outcomes," *RAND Journal of Economics*, Vol. 33, No. 3, pp. 399–433.

Becker, Gary S and Kevin M Murphy (1993) "A simple theory of advertising as a good or bad," *The Quarterly Journal of Economics*, Vol. 108, No. 4, pp. 941–964.

Bundorf, Kate, Maria Polyakova, and Ming Tai-Seale (2019) "How do humans interact with algorithms? experimental evidence from health insurance,"Technical report, National Bureau of Economic Research.

Chen, Daniel L, Tobias J Moskowitz, and Kelly Shue (2016) "Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires," *The Quarterly Journal of Economics*, Vol. 131, No. 3, pp. 1181–1242.

Dai, Weijia Daisy, Ginger Jin, Jungmin Lee, and Michael Luca (2018) "Aggregation of consumer ratings: an application to Yelp. com," *Quantitative Marketing and Economics*, Vol. 16, No. 3, pp. 289–339.

DellaVigna, Stefano and Matthew Gentzkow (2010) "Persuasion: empirical evidence," *Annu. Rev. Econ.*, Vol. 2, No. 1, pp. 643–669.

DellaVigna, Stefano and Ethan Kaplan (2007) "The Fox News effect: Media bias and voting," *The Quarterly Journal of Economics*, Vol. 122, No. 3, pp. 1187–1234.

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey (2015) "Algorithm aversion: People erroneously avoid algorithms after seeing them err.," *Journal of Experimental Psychology: General*, Vol. 144, No. 1, p. 114.

——— (2018) "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them," *Management Science*, Vol. 64, No. 3, pp. 1155–1170.

Dubois, Pierre and Tuba Tunçel (2021) "Identifying the effects of scientific information and recommendations on physicians' prescribing behavior," *Journal of Health Economics*, Vol. 78, p. 102461.

Ebenstein, Avraham, Victor Lavy, and Sefi Roth (2016) "The long-run economic consequences of high-stakes examinations: Evidence from transitory variation in pollution," *American Economic Journal: Applied Economics*, Vol. 8, No. 4, pp. 36–65.

Epstein, Andrew J and Jonathan D Ketcham (2014) "Information technology and agency in physicians' prescribing decisions," *The RAND Journal of Economics*, Vol. 45, No. 2, pp. 422–448.

Eren, Ozkan and Naci Mocan (2018) "Emotional judges and unlucky juveniles," *American Economic Journal: Applied Economics*, Vol. 10, No. 3, pp. 171–205.

Ganju, Kartik K, Hilal Atasoy, Jeffery McCullough, and Brad Greenwood (2020) "The role of decision support systems in attenuating racial biases in healthcare delivery," *Management Science*, Vol. 66, No. 11, pp. 5171–5181.

Gentzkow, Matthew and Jesse M Shapiro (2006) "Media bias and reputation," *Journal of Political Economy*, Vol. 114, No. 2, pp. 280–316.

Hallinan, Blake and Ted Striphas (2016) "Recommended for you: The Netflix Prize and the production of algorithmic culture," *New media & society*, Vol. 18, No. 1, pp. 117–137.

Hanna, Rema N and Leigh L Linden (2012) "Discrimination in grading," *American Economic Journal: Economic Policy*, Vol. 4, No. 4, pp. 146–68.

Hastings, Justine S, Mark Howison, and Sarah E Inman (2020) "Predicting high-risk opioid prescriptions before they are given," *Proceedings of the National Academy of Sciences*, Vol. 117, No. 4, pp. 1917–1923.

Heyes, Anthony and Soodeh Saberian (2019) "Temperature and decisions: evidence from 207,000 court cases," *American Economic Journal: Applied Economics*, Vol. 11, No. 2, pp. 238–65.

Huang, Jiekun, Nianhang Xu, and Honghai Yu (2020) "Pollution and Performance: Do Investors Make Worse Trades on Hazy Days?" *Management Science*.

Kamenica, Emir and Matthew Gentzkow (2011) "Bayesian persuasion," *American Economic Review*, Vol. 101, No. 6, pp. 2590–2615.

Kawaguchi, Kohei (2021) "When will workers follow an algorithm? A field experiment with a retail business," *Management Science*, Vol. 67, No. 3, pp. 1670–1695.

Kuran, Timur and Scott Lustig (2012) "Judicial biases in Ottoman Istanbul: Islamic justice and its compatibility with modern economic life," *The Journal of Law and Economics*, Vol. 55, No. 3, pp. 631–666.

Lavy, Victor, Edith Sand, and Moses Shayo (2018) "Charity begins at home (and at school): Effects of religion-based discrimination in education,"Technical report, National Bureau of Economic Research.

Martin, Gregory J and Ali Yurukoglu (2017) "Bias in cable news: Persuasion and polarization," *American Economic Review*, Vol. 107, No. 9, pp. 2565–99.

McCarthy, Ian M and Michael Darden (2017) "Supply-side responses to public quality ratings: evidence from Medicare Advantage," *American Journal of Health Economics*, Vol. 3, No. 2, pp. 140–164.

McCullough, Jeffrey S, Michelle Casey, Ira Moscovice, and Shailendra Prasad (2010)

"The effect of health information technology on quality in US hospitals," *Health affairs*, Vol. 29, No. 4, pp. 647–654.

McCullough, Jeffrey S, Stephen T Parente, and Robert Town (2016) "Health information technology and patient outcomes: the role of information and labor coordination," *The RAND Journal of Economics*, Vol. 47, No. 1, pp. 207–236.

McNamara, Cici (2021) "Health information technology and innovation diffusion in primary care settings."

Miller, Amalia R and Catherine E Tucker (2011) "Can health care information technology save babies?" *Journal of Political Economy*, Vol. 119, No. 2, pp. 289–324.

Norris, Samuel (2019) "Examiner inconsistency: Evidence from refugee appeals," *University of Chicago, Becker Friedman Institute for Economics Working Paper*, No. 2018-75.

Parente, Stephen T and Jeffrey S McCullough (2009) "Health information technology and patient safety: evidence from panel data," *Health Affairs*, Vol. 28, No. 2, pp. 357–360.

Parsons, Christopher A, Johan Sulaeman, Michael C Yates, and Daniel S Hamermesh (2011) "Strike three: Discrimination, incentives, and evaluation," *American Economic Review*, Vol. 101, No. 4, pp. 1410–35.

Price, Joseph and Justin Wolfers (2010) "Racial discrimination among NBA referees," *The Quarterly Journal of Economics*, Vol. 125, No. 4, pp. 1859–1887.

Rayo, Luis and Ilya Segal (2010) "Optimal information disclosure," *Journal of Political Economy*, Vol. 118, No. 5, pp. 949–987.

Ribers, Michael A and Hannes Ullrich (2019) "Battling antibiotic resistance: can machine learning improve prescribing?".

——— (2020) "Machine Predictions and Human Decisions with Variation in Payoffs and Skill."

Shayo, Moses and Asaf Zussman (2011) "Judicial ingroup bias in the shadow of terrorism," *The Quarterly Journal of Economics*, Vol. 126, No. 3, pp. 1447–1484.

Stigler, George J and Gary S Becker (1977) "De gustibus non est disputandum," *The American Economic Review*, Vol. 67, No. 2, pp. 76–90.

Vatter, Benjamin (2021) "Quality Disclosure and Regulation: Scoring Design in Medicare Advantage."

Appendix

A1 Additional Institutional Details

Firms' premiums are risk-rated at the firm level, starting with a flat fee of 0.93 percent of the payroll plus an additional 0-6.8 percentage points based on the risk level and the accident history of each company.

Workers' compensation insurance covers medical expenses, and provides a subsidy that replaces the worker salary during the time the worker is not able to work, a compensation if the employees ability to work is reduced by 15 to 40 percent [16], a pension benefit for those who reduced their ability to work by 40 percent or more, that covers the worker's average remuneration until the retirement age, and a survival pension to the wife and children of the worker or pensioned in case of death[17][18]. On top of financial coverage, these institutions have the obligation to work in prevention measures to help them reduce accident rates and improve workplace safety.[19]

A1.1 Accidents Committee

In order to ensure the consistency of the coverage decision among doctors there is an accident management meeting (Reunión de gestión de Accidentados - RGA). The RGA is a meeting that takes place at 12:00 am everyday, in every clinic of the network. The goal of the RGA is to review all the cases from the previews day and validate the coverage decision. Three roles take part in the meeting, with various responsibilities:

- Chief doctor: reviews the event description of the patient and compares what was said to the admission executive when the patient first arrived and to the

[16]The compensation is given for one time and is equivalent to 1.5 to 15 times the worker salary

[17]The pension for the widow is 42 percent of the regular salary if she or he does not have children with the employee, and 35 percent if she or he is the mother or father of the employee's children. The mother or father of non-marital filiation receives between 21 and 25,2 percent of the death worker's salary (30 to 36 percent of the pension). Each children receives 14 percent of the regular salary (20 percent of the pension)

[18]For more information about monetary benefits: https://www.suseso.cl/613/w3-propertyvalue-63795.html [April 18th, 2021]

[19]Paradoxically, this puts direct pressure on the mutuals earning potential, reducing income as they are successful reducing accident rates.

doctor (to identify if there are inconsistencies). Additionally the chief doctor analyzes the clinical information of the patient and reevaluates the coverage decision made by the doctor ”tratante”

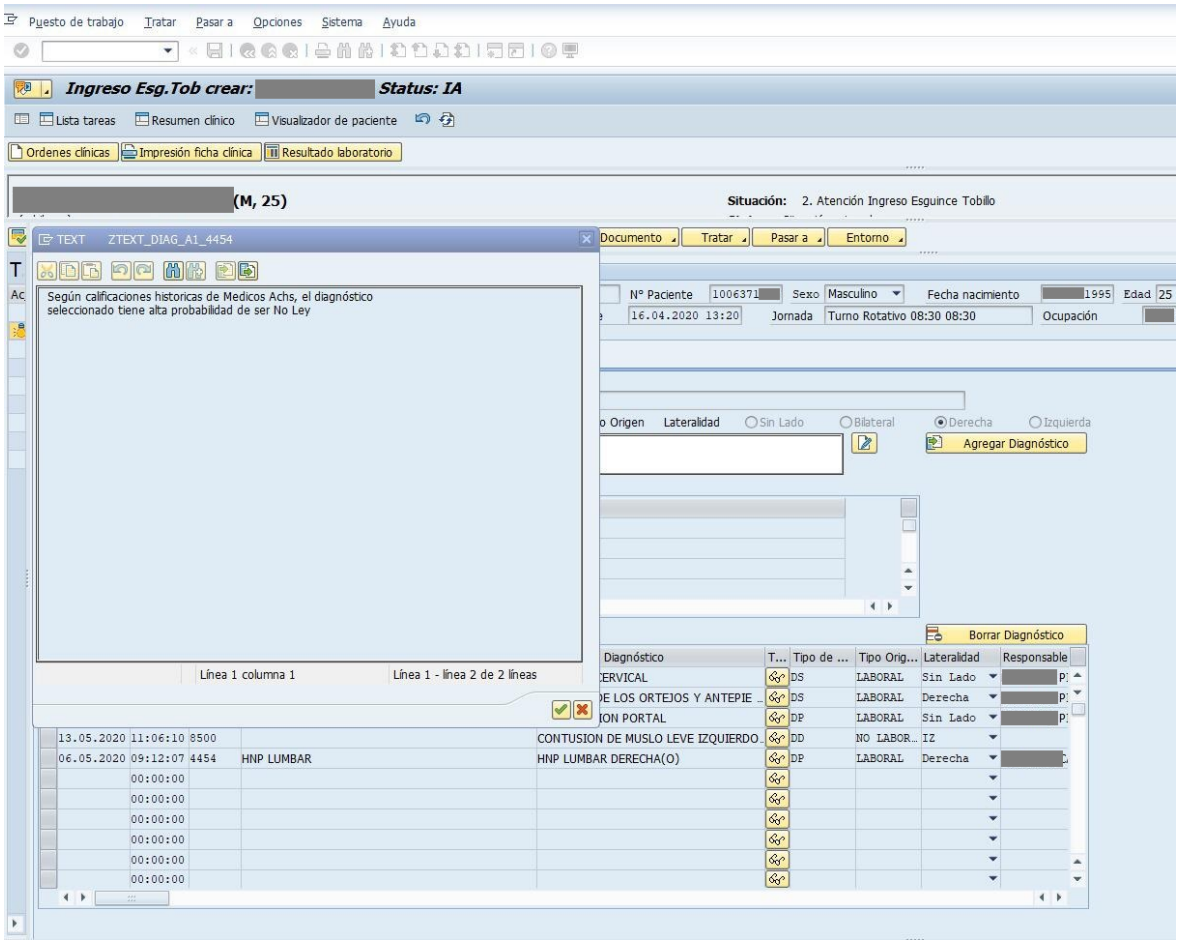
- ”Jefe de Gestión Comercial y Servicios Preventivos - JGCySP”: makes sure that the event (siniestro) belongs to that clinic and identifies anomalies related to the accident (e.g. the patient indicates that he was moving a 100kg box and the JGCySP identifies that there are boxes of that size/weight in the workers company). When these inconsistencies are identified the committee requests a report to the workplace safety consultant/expert, that goes to the company site and validates the event description provided by the patient. This report is used to define the final coverage status of the accident

- ”Rol Calificador”: The person responsible for defining administrative coverage based on what has been discussed in the RGA. In order to revoke coverage because of administrative issues it is mandatory that the ”Rol Calificador” provides the report done by the workplace safety consultant. Because of this requirement the change in coverage for administrative reasons is done by the ”Rol Calificador” after the RGA

A2 Alert Screenshot

Figure A1 shows the physician’s computer interface and the pop-up alert.

Figure A1: Alert Screenshot



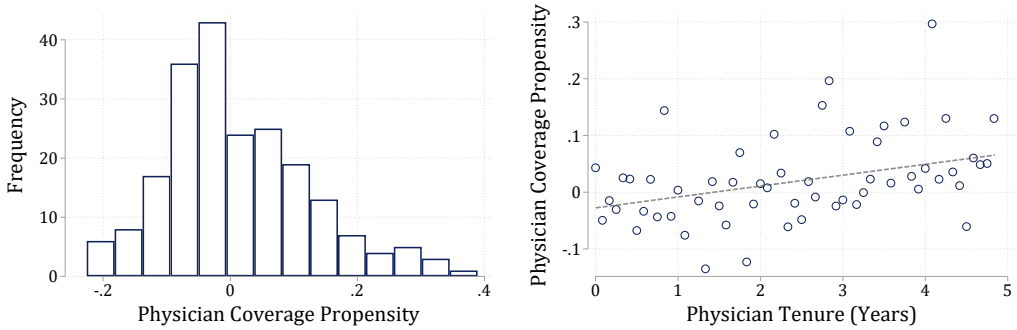
A3 Additional Tables and Figures

Table A1: Correction Probability. Heterogeneous effects by Physician Characteristics

	(1)	(2)	(3)	(4)	(5)
Treated * Post	0.011*	0.012**	0.011*	0.012*	0.012*
	0.006	0.006	0.006	0.007	0.007
Treated * Post * Alert	-0.020***	-0.016**	-0.022***	-0.017***	-0.017***
	0.006	0.007	0.006	0.006	0.006
Treated * Post * Alert * Tenure	-0.003*				-0.002
	0.001				0.003
Treated * Post * Alert * ln(Visits)		-0.006			
		0.004			
Treated * Post * Alert * ln(Visits-Diag)			-0.007**		
			0.003		
Treated * Post * Alert * Generosity				-0.013***	-0.013***
				0.002	0.003
Treated * Post * Alert * Generosity * Tenure					0.001
					0.001
N	62315	62315	62315	62245	62245
No. Clusters	79	79	79	79	79
79					
R-Squared	0.03	0.03	0.03	0.03	0.03

Note: The table shows the results main specification where the DiD coefficients are interacted with physician characteristics. Contested sample. Only physicians with more than 30 visits in pre-period. All specifications include diagnosis group, time, patient characteristics, and physician fixed effects; and month×year and day-of-the-week fixed effects. Bootstrapped standard errors in Columns (4)-(5). Standard errors clustered at the clinic level * p<0.10, ** p<0.05, *** p<0.01

Figure A2: Physician Coverage Propensity



(a) Physician **(b)** Interaction with Tenure

Note: Panel (a) shows the distribution of physicians coverage propensity, where coverage propensities are the physician fixed effects of physicians' coverage decisions after controlling for diagnosis group, month×year, and day-of-the-week fixed effects; and patient characteristics (quadratic polynomial on patient's age, and patient's gender and employer's 1-digit SIC fixed effect). Panel (b) shows the correlation between coverage propensities and physician tenure.

A4 Balanced Sample Tests

Table A2: Logit Regressions

	Dep. Var.: Treatment Clinic							
	Weekly Mean				Weekly Median			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Coverage	-0.136	-0.227			-0.093	-0.125		
	(0.181)	(0.345)			(0.122)	(0.223)		
Δ Coverage		0.144				0.074		
		(0.175)				(0.113)		
Correction			0.359	0.653			0.121	-0.056
			(0.231)	(0.514)			(0.265)	(0.573)
Δ Correction				-0.432				-0.112
				(0.265)				(0.303)
N	11344	9716	11344	9716	11344	9716	11344	9716
No. Physicians	218	213	218	213	218	213	218	213
Pseudo R-Sq.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note: The table shows the results of logit regressions where the dependent variable is the selection into treatment specification. All specifications include time (month×year) fixed effects. An observation is a physician-week unit. Standard errors clustered at the physician level.