Decision-Making with Machine Prediction: Evidence from Predictive Maintenance in Trucking[*]

Adam Harris[†] Maggie Yellen[‡]

Job Market Paper

October 23, 2023

Click here for the latest version.

Abstract

In this paper, we study the role of predictive artificial intelligence (AI) in human decision-making. Using a rich decision-level data set from the maintenance of heavy-duty trucks, we document how the repair decision-making of expert technicians changes with the introduction of an AI tool designed to predict the risk of truck breakdowns. We develop and estimate a dynamic discrete choice model of technician decision-making. The resulting estimates show that technicians with the AI tool exhibit a substantially better ability to predict breakdown risk than those without the tool. This improvement in predictive ability translates into better outcomes: The AI tool reduces the total costs that technicians incur by $343-$686 per truck per year. Furthermore, with the AI tool, technician decision-making is nearly optimal; only 5% more cost savings could feasibly be achieved with further improvements in decision-making quality.

[*] We are very grateful to Daron Acemoglu, Nikhil Agarwal, Nancy Rose, Tobias Salz, and Mike Whinston, all of whom have been so generous with their guidance and support. We also thank the many people in the trucking industry who shared their time and knowledge with us. We are particularly grateful to the people within the firm whose data we use in this paper both for sharing their data with us and for taking the time to help us understand their maintenance operations. We also thank all participants in the MIT Industrial Organization Seminar, the MIT Industrial Organization Lunch and the MIT Organizational Economics Lunch for their comments and feedback. We acknowledge the support of the George and Obie Shultz Fund. This material is also based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1745302. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Finally, Harris acknowledges support from the US Department of Transportation under NSF grants SES1559013 and SES2315269 to the NBER. Any views expressed are those of the authors and should not be interpreted as representing the views of the USDOT or NBER.

[†] Corresponding author: MIT Department of Economics and NBER, asharris@mit.edu.

[‡] MIT Department of Economics, myellen@mit.edu.

1 Introduction

Every day, professional human decision-makers throughout society face choices that determine socially and economically important outcomes. Doctors choose treatment plans for their patients; bankers choose whether to approve or deny loan applications; and, in various industrial settings, technicians choose whether or not to repair machines. The quality of these professionals' decision-making depends both on their ability to make *predictions*—about, for instance, whether the patient has a tumor, whether the borrower will default, or whether the machine will break down—and on their ability to make *judgments*—combining these predictions with their evaluations of payoffs to choose an action. Increasingly, however, humans do not navigate this decision-making process alone: in a variety of settings, artificial intelligence (AI) tools now assist professional human decision-makers with prediction.

AI is undoubtedly a powerful tool for prediction. Indeed, in many domains, algorithms have achieved predictive ability superior to that of expert humans.[1] Nevertheless, it is a priori unclear whether providing humans with AI predictions necessarily improves the quality of decision-making; this is ultimately determined by how humans use—or misuse—these AI predictions.[2]

In this paper, we study the role of a high-quality predictive AI tool in the decision-making of technicians charged with the maintenance of heavy-duty trucks.[3] Observing the engine repair decisions that technicians make before and after the introduction of the AI tool, we explore how technicians use the tool. We quantify its value in terms of improvements (if any) in the quality of technicians' repair decisions. To accomplish this, we estimate a dynamic structural model of technician decision-making. By exploiting comprehensive truck-generated data, we are able to separately identify and estimate costs and technicians' beliefs about breakdown risk. Our estimates show that with the AI tool, technicians' beliefs more accurately reflect the true risk of breakdown. This results in an improvement in technician decision-making, leading to a substantial ($343-$686) reduction in annual per-truck maintenance expenditures.

This analysis makes use of data from a large private fleet of heavy-duty trucks, which, in early 2020, introduced an AI tool—which we refer to as *PredictFix*—that provides technicians with algorithmically-generated alerts indicating when a truck has a high risk of breaking down. We observe not only technicians' engine repair decisions but also a rich set of truck-generated

[1] To cite just a few examples, Jalalifar et al. (2022) develop an algorithm that outperforms the ability of oncologists to predict the effects of radiology on brain tumors; Irvin et al. (2019) train a model that predicts the presence of chest pathologies (e.g., pneumonia) better than radiologists; Van Binsbergen et al. (2023) train a model that forecasts equity prices more accurately than professional analysts.

[2] Indeed, some evidence lab experiments, including that of Agarwal et al. (2023), shows that assistance by a high-quality AI tool does not, on average, improve the quality of decision-making.

[3] Heavy-duty trucks are sometimes also called tractor-trailers, semi trucks, or eighteen-wheelers.

data, including all of the sensor measurements and fault codes available to technicians. Taken together, these signals give us a comprehensive and high-dimensional description of the state of every truck at every point in time. Section 3 provides a detailed description of these data, as well as the PredictFix AI tool.

We begin our empirical analysis in Section 4 by presenting descriptive evidence on the PredictFix AI tool. First, we show that PredictFix is a high-quality predictor of breakdowns. Next, we show that technicians' engine repair decisions are moderately responsive to the alerts that PredictFix generates: a PredictFix alert increases the probability of a same-week engine repair by 15.0pp (for reference, the unconditional weekly repair probability is 10.5%). These two facts would seem to suggest that PredictFix improves decision-making quality, and therefore outcomes; however, comparing outcomes for the "pre period" and the "post period," we find no change in the frequency of either repairs or breakdowns. This suggests that the effects of PredictFix may be confounded by changes in other payoff-relevant variables that coincided with PredictFix's introduction. Consistent with this hypothesis, we present suggestive evidence that technicians faced higher repair costs in the post period than in the pre period.

To quantify the value of PredictFix while controlling for costs, in Section 5, we develop a dynamic structural model of technician decision-making. At the heart of the model is a simple trade-off between the technician's perceived risk of a breakdown and the cost of doing a repair. By taking this model to the data, we can estimate both the technician's beliefs about breakdown risk and their costs, both before and after PredictFix's introduction. In our setting, separate identification of these beliefs and costs is made possible by the richness of our comprehensive, high-dimensional truck-generated data. Yet this high-dimensional state also presents a challenge for estimation, as standard techniques for estimating dynamic models are ill-suited to such high dimensionality. To overcome this challenge, we develop a novel approach that builds on the insights of Hotz and Miller (1993) and Arcidiacono and Miller (2011). Section 6 describes how we carry out this approach in practice to recover estimates of costs and the technician's beliefs about breakdown risk.

Section 7 compares our estimates of beliefs and costs with versus without PredictFix. First, to understand how PredictFix affects technicians' beliefs about breakdown risk, we evaluate the quality of technicians' perceived risk of breakdown—both with and without PredictFix—as a predictor of actual breakdowns. We find that, with PredictFix, technicians exhibit a substantially better ability to predict breakdowns (AUC-ROC is 0.704 with PredictFix versus 0.598 without PredictFix).[4] Second, we present our costs estimates, which show that both the mean

[4] AUC-ROC, which refers to the area under the receiver operating characteristic (ROC) curve, is a quantitative measure of the quality of a continuous predictor of binary outcomes. A value closer to 1 indicates a higher-quality predictor. See Section 4.1 and Appendix B.1 for a discussion of ROC curves.

and the variance of repair costs are higher in the post period as compared with the pre period.

Finally, in Section 8, we use the estimated dynamic model to evaluate a counterfactual in which we simulate introducing PredictFix while holding cost conditions fixed. This enables us to quantify the *value of PredictFix* in terms of the change in total expenditures that results from PredictFix's effect on technician decision-making. We find evidence of a substantial improvement: technicians with PredictFix achieve a reduction in total engine maintenance costs of $343 - $686 per truck per year relative to those without PredictFix. This represents 95% of all the cost savings that could feasibly be achieved with improved repair decision-making, meaning that with PredictFix, technician decision-making is very nearly optimal.

The paper proceeds as follows: Section 2 summarizes the related literature and our contribution. Section 3 describes our setting—the maintenance of heavy-duty trucks—and the data we use to analyze technicians' use of PredictFix in this setting. Section 4 presents descriptive evidence in the form of five key facts, which highlight critical features of the setting and begin to shed light on technicians' use of PredictFix. Section 5 develops a dynamic structural model of technician decision-making suitable for empirical analysis. Section 6 describes how we estimate this model. Section 7 presents key estimates, shedding light both on costs and on technicians' ability to predict breakdowns with and without PredictFix. Section 8 presents the results of the counterfactual analysis that speaks to the value of PredictFix. Finally, Section 9 concludes the paper.

2 Literature review

Our paper primarily contributes to the emerging empirical literature on predictive algorithms as inputs to decision-making processes.[5] This literature comprises both experimental and observational studies. Experimental work within economics includes Agarwal et al. (2023), who find that providing radiologists with AI prediction does not improve the quality of their diagnoses on average. Experimental work in the medical and computer science literature includes Tschandl et al. (2020) (skin cancer), Kim et al. (2020) (breast cancer), Reverberi et al. (2022) (colon cancer), and Jakubik et al. (2022) (loan approval decisions). A smaller number of papers use observational data from settings where predictive algorithms were actually implemented to study how human decision-making changes in the presence of algorithmic prediction. Within economics, Stevenson and Doleac (2022) and Angelova et al. (2022) study how judges use risk assessment tools in making sentencing and pretrial detention decisions.

[5] There is also a closely related and more extensive empirical economics literature on decision-making under uncertainty in which algorithms are not the focus. Papers in this literature include Mullainathan and Obermeyer (2022), Currie and MacLeod (2020), Abaluck et al. (2020), and Chandra and Staiger (2007).

Neither paper finds evidence that the risk assessment tool led to improvements in relevant outcomes, such as public safety and the incarceration rate. Related, though less directly relevant to our paper, Albright (2023) studies the effect of algorithmic *recommendations* on judge's bail decisions.[6] Outside of economics, observational studies on this topic include De-Arteaga et al. (2020), who study how child welfare workers use a predictive algorithm in their screening decisions.

Relative to this existing work, we make three contributions. *First*, we estimate the dollar-denominated *value* of a predictive AI tool by quantifying the social welfare gain (or loss) resulting from the effects of the AI tool on the quality of technician decisions. This is made possible both by our setting—placing a dollar value on social welfare effects is more straightforward in truck maintenance than in, for instance, the judicial context—and by our use of a decision framework to quantify the costs and benefits of technician decisions. *Second*, we study how AI prediction changes human decision-making in a context where payoffs are inherently dynamic. In our setting, a technician makes a repair decision for a truck not only this week, but also next week, the following week, etc. In contrast, previous studies have focused on static settings where, for instance, a judge or radiologist has a one-off interaction with a defendant or patient. To account for dynamics in our quantification of the effects of the predictive AI tool, we develop and estimate a dynamic model of technician decision-making. *Finally*, we study the use of predictive AI in a new and economically significant setting: the maintenance of heavy-duty trucks, which play a central role in the US goods economy. Although the settings of previous research (e.g. criminal justice and medicine) are undeniably of great social importance, our paper provides insight into the use of AI tools in the industrial settings in which they are increasingly being adopted.[7]

Methodologically, our paper draws upon and contributes to the literature on estimating dynamic discrete choice (DDC) models. Building on the insights of Rust (1987), Hotz and Miller (1993), and Arcidiacono and Miller (2011), we describe and demonstrate an approach to estimating DDC models in high-dimensional settings. Our approach obviates the need to flexibly estimate the transition process, which is generally infeasible for a high-dimensional state variable. Instead, we employ a distributional assumption under which the dynamic component of payoffs can be expressed as a closed-form function of objects that are feasibly estimated from high-dimensional data.

[6] PredictFix, the AI tool we study, provides technicians with algorithmic *predictions*, rather than recommendations. In our setting, payoffs depend on both breakdown risk and costs. PredictFix provides technicians with a signal about breakdown risk, but in no way incorporates information about costs. It is up to the technician to combine PredictFix's output with her evaluation of costs to arrive at a repair decision.

[7] Zolas et al. (2021) note that, in the 2018 Annual Business Survey, the four-digit NAICS code industries reporting the highest level of adoption of machine learning are "Metalworking Machinery Manufacturing" and "Machine Shops; Turned Products; Screw, Nut and Bolt Manufacturing."

We also contribute to an empirical literature on the trucking industry. Most closely related are papers that study the introduction of new technologies in trucking (e.g., Hubbard (2000, 2003); Baker and Hubbard (2003, 2004); Yang (2022); Armitage et al. (2023)) and those that study the maintenance of fleet vehicles (i.e., Rust (1987)). Other economics papers related to trucking include Rose (1985, 1987), Hubbard (2001), and Harris and Nguyen (2021, 2022).

Finally, our paper can be thought of as continuing a long industrial organization tradition of single-firm productivity case studies. The goal of such studies is to estimate productivity over time and, in some cases, to explore the determinants of productivity's evolution. Examples include Gort and Sung (1999) and Hendel and Spiegel (2014). Our analysis of the effect of a predictive AI tool on the quality of technician decision-making is analogous to estimating the productivity of the fleet's engine maintenance operation before and after the introduction of the tool.

3 Heavy-duty truck maintenance: Setting and data

We study the engine repair decisions of technicians charged with maintaining a fleet of heavy-duty trucks. This setting is particularly well suited to studying how skilled human decision-makers use a predictive AI tool. Like equipment in many modern industrial settings, heavy-duty trucks generate high-quality, high-dimensional data as a byproduct of regular operations. These abundant data, which contain valuable information about the risk of breakdown, make it feasible to train and use an AI tool to generate algorithmic breakdown predictions. We observe repair decisions and truck-generated data from a large private fleet that adopted such a tool.

3.1 Setting: The fleet, the technician's problem, and the algorithm

Trucks play a central role in the US goods economy, transporting more than 72% of all domestic freight by volume.[8] Fulfilling this role, however, requires that trucks be kept in good health; failure of any of the thousands of components of a truck can result in a costly breakdown. We study engine repair decisions made by several hundred technicians at a large private fleet. The fleet is owned and operated by a firm—which we will refer to using the pseudonym *PFC* (private fleet company)—that manufactures and distributes consumer goods across the US. We focus in particular on the fleet's heavy-duty trucks.

In making repair decisions, technicians are informed by rich truck-generated data. A modern heavy-duty truck is equipped with as many as several hundred sensors, each of which

[8] Bureau of Transportation Statistics Freight Facts and Figures 2017, Department of Transportation.

measures the performance of some particular truck component. These sensors, in conjunction with an on-board computer, generate a high-frequency stream of sensor measurements and fault codes, which PFC makes available to its technicians. Prior to March 2020, PFC's technicians were expected to—without any algorithmic assistance—use this truck-generated data to form breakdown-risk predictions and make engine repair decisions. This is a challenging task, both because trucks are complex machines and because the set of truck-generated data is so vast; in a typical day, a heavy-duty truck produces about 4,000 sensor measurements and 10 fault codes, making constant human monitoring of all of these signals infeasible.[9]

To assist technicians with this task, in March 2020, PFC purchased an AI breakdown-prediction tool, which we will refer to as *PredictFix*, from a technology firm specializing in industrial machine learning prediction. This tool takes as input a truck's sensor measurements and fault codes and, using machine learning models, generates an *alert* when the predicted risk of failure for a component—something likely to result in an engine breakdown—is high. For example, there are PredictFix alerts categories that indicate a high risk of engine overheating, cylinder head failure, and coolant leak. Each alert comes with an associated "severity" level, either high-severity or medium-severity.[10] Note that, since PredictFix alerts are deterministic functions of sensor and fault data that is also available to technicians, PredictFix does *not* provide technicians with new information on the state of the truck; rather, it provides the same information in a form that is potentially easier for technicians to interpret and make use of.

We think of PFC's objective as minimizing a (broadly defined) notion of costs. The costs of repairs and breakdowns include both tangible and intangible components. With either a repair or breakdown, tangible costs include labor and materials; when a breakdown occurs, tangible costs may also include towing costs and the cost of leasing at temporary replacement truck. Intangible costs include both the opportunity cost of a truck not being on the road, as well as the shadow cost of constraints imposed by each PFC facility's capacity to do repairs. In addition, when a breakdown occurs, PFC incurs potentially large disruption costs. PFC's fleet management team views relationships with both retailers and drivers as very valuable. The firm maintains explicit and implicit service-level agreements with retailers; failure to deliver product on time due to a breakdown may damage these valuable relationships. Moreover, breakdowns can cause substantial inconvenience for drivers. PFC views the market for such talent as highly competitive, so retaining drivers is a high-priority objective. For these reasons, such disruption costs often exceed the tangible, monetary costs of a breakdown. For a discussion of technicians'

[9] A typical technician is responsible for 30-60 trucks.

[10] Note that these severity level labels are not an output of the algorithm. Rather, the PredictFix fleet management team assigned these labels to each of the PredictFix alert categories. For instance, of the three alert categories listed above, they labeled engine overheating and cylinder head failure as high-severity and labeled coolant leak as medium-severity.

incentives and possible agency issues, see Appendix A.

3.2 Data

To understand technician decision-making, we study the relationship between the truck- and AI-generated data that technicians observe and the engine repair decisions they make. This analysis requires data on sensor measurements, fault codes, and PredictFix alerts, as well as data on engine repair and breakdown events.

We obtain truck-generated sensor and fault data from the portal that technicians use to view sensor measurements and fault codes. We thus observe all of the quantitative data in the technician's information set, which provides a comprehensive description of the state of each truck at each point in time. Our analysis uses this sensor and fault data for two disjoint time periods. The first, September 2019 through March 2020, covers the six months immediately before the introduction of PredictFix—the *pre period* for our analysis. The second, March 2021 through November 2022, covers the twenty months starting one year after the introduction of PredictFix—our *post period*.[11]

Figures 1 and 2 give a sense of the high frequency and enormous scale of the truck-generated data available to technicians. Figure 1 shows the pattern of measurements for a selected set of sensors on one truck over a 24-hour period. As this example illustrates, sensor measurements are high-frequency and track component performance (e.g. engine cranking voltage, engine manifold pressure), driving patterns (e.g., speed, acceleration, braking) and environmental conditions (e.g. outside air temperature and pressure). Panel (a) of Figure 2 shows the distribution of the number of sensor measurements per truck-month; the median truck-month has about 100,000 sensor readings.

Vehicles' on-board computers are factory-programmed with a set of several thousand simple rules, each of which translates patterns of particular (and potentially concerning) sensor measurements into "fault codes." For example, an "Engine misfire for multiple cylinders" fault code might be triggered by measurements consistent with a fuel injector or spark plug issue; a "Gas supply pressure—Data valid but below normal operational range" fault, meanwhile, is triggered by one or more low pressure measurements and might indicate any number of issues with the fuel system. Figure 2 (b) shows the distribution of the number of faults per truck-month; the median truck-month has 136 fault codes, though the distribution has a long right tail.[12] While the occurrence of a fault does sometimes indicate a serious issue, the rate of false

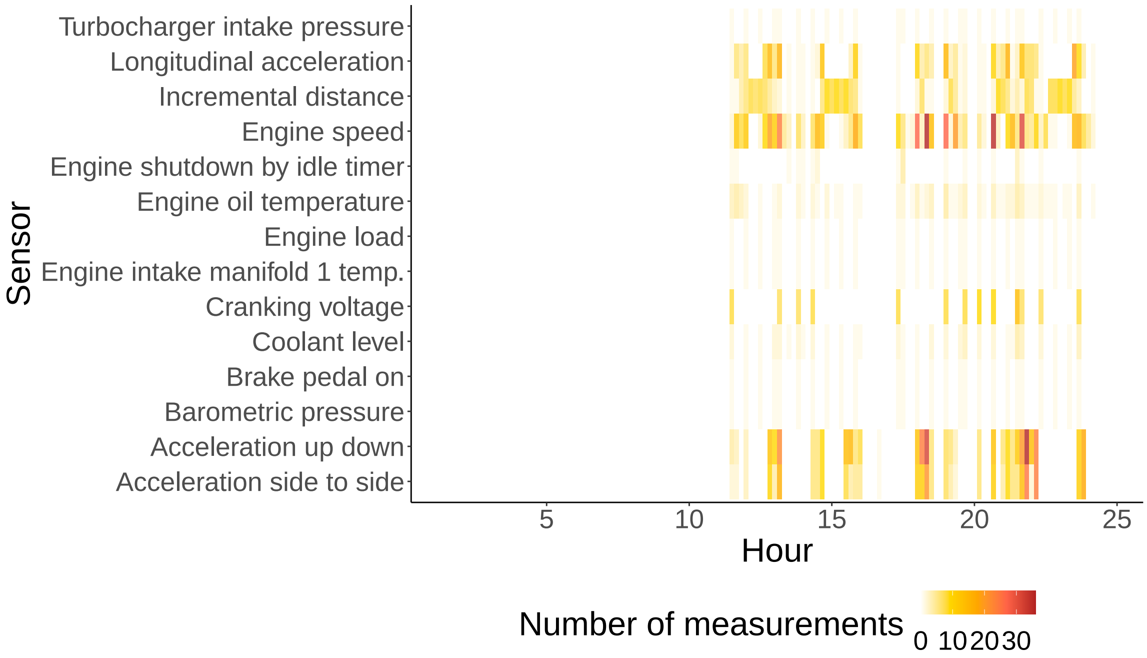
[11] The range of dates was dictated by a key constraint: Due to the enormous volume of fault and sensor data, PFC does not generally retain more than six months of these data at a time. Fortunately, PFC's telematics software provider was able to restore a six-month snapshot of older data, corresponding to our pre period.

[12] 21.6% of truck-months have more than 500 fault codes; 10.0% have more than 1000.

positives is, according to technicians, extremely high. Even if it were possible for technicians to respond to every fault, doing so would not be desirable, as the vast majority are not indicative of a genuine issue requiring technician intervention.

Our analysis also uses data on PredictFix alerts. For each alert that PredictFix generates, we observe the truck identifier, the timing of the alert, the component-specific model that triggered the alert, and the corresponding severity level. Figure 2 (c) illustrates the frequency of these alerts; of all truck-months after March 2021, about 40% have at least one PredictFix alert.

Figure 1: Sensor measurements recorded over the course of a day: An example

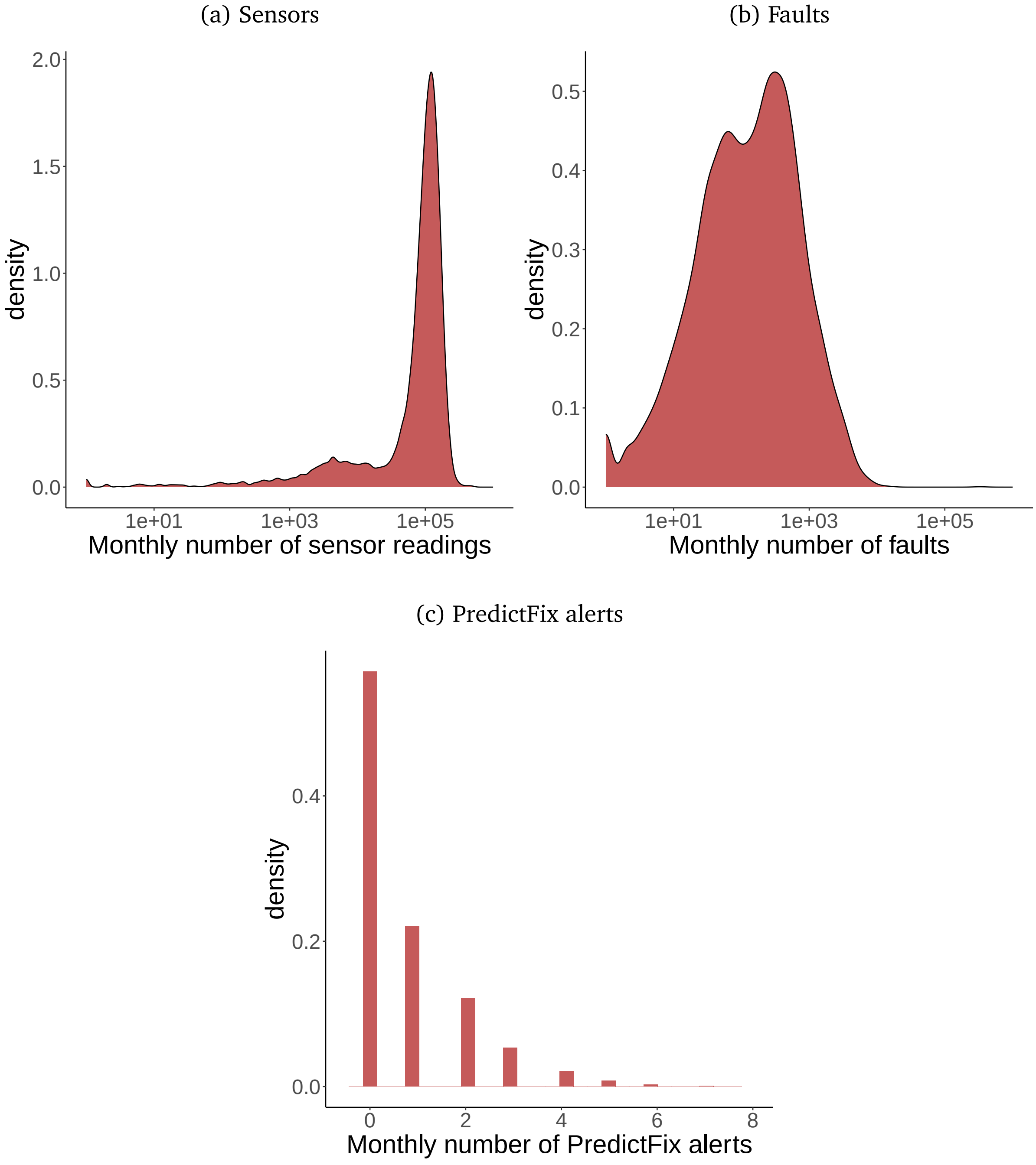


Notes: This figure illustrates the incidence of recorded sensor measurements for one truck over the course of a day. The y-axis lists a subset of fourteen onboard sensors; for each sensor type and each ten-minute interval, the color of the corresponding cell indicates the number of sensor measurements recorded. Based on the pattern of these sensor readings, we can see that the truck was actively in use from about 11 AM until midnight and that readings for each sensor were recorded periodically throughout that period.

Taken together, the sensor measurements, fault codes, and PredictFix alerts give technicians—and us—a rich set of data on the state of the fleet's trucks. For the periods described above and the set of about 700 heavy-duty trucks used in our analysis, we observe about 1.3 billion sensor measurements, 8.3 million fault codes, and nearly 3,000 PredictFix alerts.

Finally, we make use of PFC's fleet maintenance records. These records are both comprehensive—they describe every maintenance event for every vehicle in the fleet—and rich in detail—for every event, they list dates, times, technician identifiers, the reason for the event, the vehicle systems and components involved, and the internal accounting costs of labor and parts

Figure 2: Number of faults, sensors, and PredictFix alerts per truck-month


(a) Sensors

(b) Faults

(c) PredictFix alerts

Notes: These histograms contrast the frequency of truck-generated fault codes and sensor readings with the frequency of algorithm-generated PredictFix alerts. Each histogram treats the unit of observation as a truck-week and shows the distribution of the number of faults/sensors/alerts per truck-week. Note that the x-axes of Panels (a) and (b) use a log scale.

for each constituent line item. In evaluating technician decision-making, we want to focus on repairs that are elective, i.e., performed at the discretion of the technician. Therefore, our empirical definition of a “repair” excludes planned preventative maintenance, which follows a pre-determined schedule.

In addition to repair decisions, the maintenance records also give us data on a key outcome: engine breakdowns. In defining breakdown events, we want to focus on those breakdowns that are plausible foreseeable and preventable by technician intervention. We therefore treat an event as a breakdown only if it occurred because of the failure of truck components; we exclude events caused by external factors outside a technician’s control, like collisions and winter weather conditions (e.g., ice-related issues).

Preparing our extremely granular continuous-time signal, alert, and repair data for analysis requires discretizing time. To that end, we aggregate measurements for each sensor and each fault code into a set of truck-week level statistics: for each sensor type, we compute the weekly total number of measurements, the weekly maximum measurement, the weekly minimum measurement, the weekly mean measurement, and the weekly standard deviation of measurements. For each of the 334 most common fault codes, we count the number of times each fault code occurs in a week.Table 1 presents the set of variables—a combination of truck-generated data and truck history data—that, for the purpose of our analysis, comprises the technician’s information set when predicting breakdowns. The model in 5, which refers to this set of breakdown-predictor variables as x, clarifies the role that breakdown predictions play in technician decision-making.

Table 1: Variables relevant to breakdown prediction (x)

Maintenance history:	Miles/weeks since last repair
Truck characteristics:	Model year
For each of 334 fault codes:	Weekly indicator Three weeks of lagged indicators
Total faults:	Total number of weekly faults
For each of 48 sensors:	Weekly min, max, mean, std. dev., count One week lag for each

Notes: This table outlines the set of variables constructed from the data that we think of as predictors of breakdown risk. These variables correspond to x in our model (presented in Section 5).

Of the data-cleaning steps we perform, two are particularly important, as each reduces our sample size. First, we must deal with the fact that data for certain trucks is intermittently missing. Because this missing data is most often the result of PFC’s historical data storage practices, we drop all missing truck-weeks from our analysis. Second, while most repairs are completed within a few days, a repair can occasionally stretch on as a series of work orders completed over as long as several weeks. To deal with this issue, we collapse each such work order series into a single repair or breakdown event occurring at the beginning of the series. Moreover, we drop from our analysis the three weeks following every repair or breakdown event. Table 2 describes the cleaned sample.

Table 2: Sample size

	Pre	Post
Number of trucks	530	693
Number of weeks	21	81
Truck-weeks	4,406	16,925

Notes: This table presents describes the data sample used in our analysis. The two columns represent, respecitively, the pre period (before the implementation of PredictFix) and the post period (after the implementation of PredictFix).

4 Descriptive evidence: Five facts

We begin our empirical analysis by presenting five key facts. The first two facts clarify the nature of breakdown risk and the quality of PredictFix. The third and fourth facts quantify technicians' responses to PredictFix and changes in aggregate fleet outcomes. The fifth fact sheds light on changes in cost conditions around the time of PredictFix's introduction.

4.1 Fact 1: Breakdown risk is predictable.

Before studying the quality of technician predictions and decisions, we quantify the extent to which the data in the technician's information set predicts breakdowns. To accomplish this, we train a gradient-boosted decision tree (GBDT) model (Friedman, 2001)—a flexible machine learning model that combines an ensemble of decision trees—to predict engine breakdowns based on observable truck states.[13] We use x to denote sensor, fault, and maintenance history data and use $\hat{\pi}(x)$ to denote the GBDT model's prediction of breakdown risk.

To evaluate the quality of our estimated $\hat{\pi}$ as a predictor of breakdowns, we use a common tool for evaluating machine learning predictions: the Receiver Operator Characteristic (ROC) curve. The ROC curve is a production possibilities frontier (PPF) in the false positive rate (FPR)-true positive rate (TPR) space. In our context, the true positive rate is the proportion of actual breakdowns that are correctly predicted to be breakdowns, and the false positive rate is the proportion of actual non-breakdowns that are incorrectly predicted to be breakdowns.

[13] For details of the cross-validation and training of this model, see Appendix D.6.

The ROC curve illustrates the set of (FPR, TPR) pairs that are achievable with binary classifiers constructed from $\hat{\pi}(x)$. Naturally, a perfect (oracle) predictor would have (FPR,TPR) = (0,1); a completely uninformative predictor, meanwhile, would have an ROC curve on the 45-degree line. Thus, an ROC curve further above the 45-degree line and closer to the (0,1) point indicates a higher-quality predictor. For this reason, the *area under the ROC curve* (AUC or AUC-ROC) is commonly used to quantify predictive quality, with AUC=0.5 indicating a completely uninformative predictor and AUC=1 indicating an oracle predictor. (For a more thorough primer on ROC curves, see Appendix B.1.)

Figure 3 presents the out-of-sample ROC curve for our breakdown predictor $\hat{\pi}(x)$.[14] Its AUC, 0.778, indicates that $\hat{\pi}(x)$, and therefore the state of the truck x from which it is constructed, contains substantial information about breakdown risk.[15] Since x is observable to technicians, this finding indicates that a technician (at least one with unlimited attention and computational ability) could extract valuable information from x to inform her repair decision. Whether actual technicians are able to do this—both on their own and when they are assisted by PredictFix—is the central question of this paper, one that we address using estimates from our structural model in Section 7.

4.2 Fact 2: PredictFix is a good predictor of breakdown risk.

Having established that breakdown risk is predictable, we next verify that PredictFix alerts are indeed good predictors of breakdown risk. To do this, we build upon the analysis from the previous subsection: In addition to the ROC curve, Figure 3 also plots the false positive and true positive rates for PredictFix alerts, which are binary predictors of breakdowns. As $\hat{\pi}(x)$ should fully exploit the information on the state of the truck x to predict breakdowns, the points of the ROC curve represent the best achievable binary predictors of breakdowns given x and given the training data that we have.[16] Since a PredictFix alert is a binary predictor that is also a function of x, the position of the PredictFix points in the (FPR, TPR) space relative to the ROC curve indicates their quality of prediction relative to this optimal benchmark.

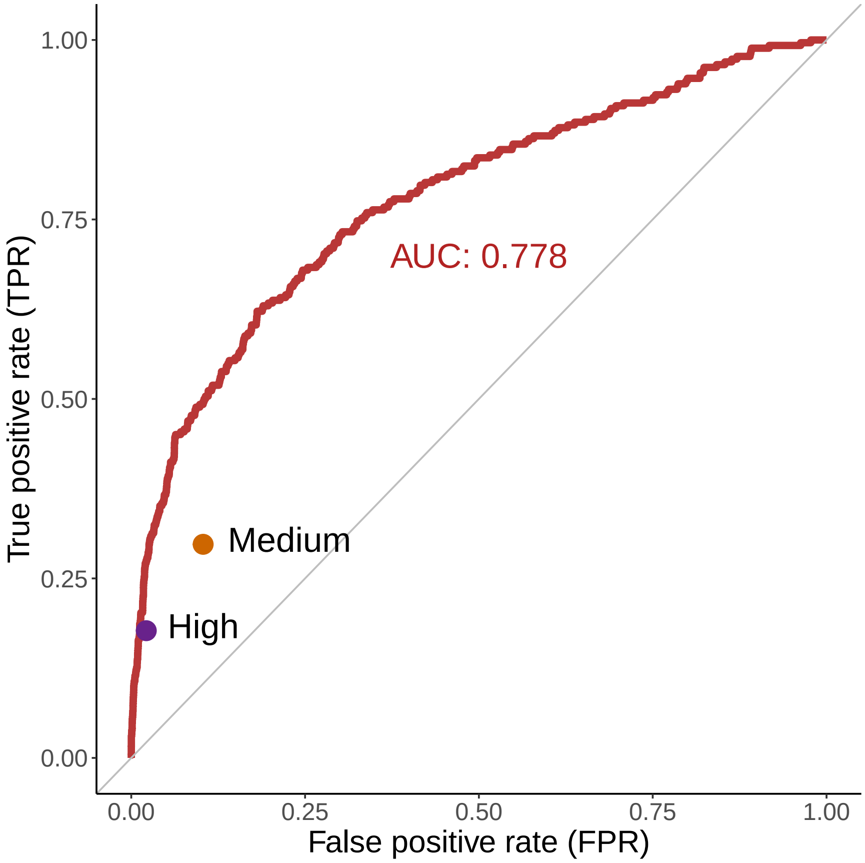
Looking at the y-coordinates (true positive rate) of the points in Figure 3, we can see that 21.4% of all breakdowns that occur in the post period are preceded by a high-severity PredictFix

[14] This is the *out-of-sample* ROC curve in the sense that the FPR and TPR values are computed using observations that were *not* used in training the GBDT model.

[15] While this predictor is far from the oracle benchmark with AUC = 1, this is not at all surprising: trucks are complex machines subjected to a wide array of human and environmental factors that are unobservable and unpredictable by technicians and by us. In other words, we intuitively understand that there is a substantial degree of *irreducible uncertainty* in the breakdown process, so no feasible predictor could achieve the oracle benchmark.

[16] This claim also assumes optimal model selection and hyperparameter tuning. Further improvements in out-of-sample fit may be possible using a different ML model (e.g., a neural network), better feature engineering, or better tuning of hyperparameters.

Figure 3: PredictFix Alerts as Predictors of Breakdowns



Notes: The ROC curve in this figure illustrates the quality of $\hat{\pi}(x)$ as a predictor of breakdowns. $\hat{\pi}(x)$ is the output of a GBDT model trained to predict breakdowns. The ROC curve is constructed using observations excluded from GBDT training, so this ROC curve captures the *out-of-sample* predictive quality of $\hat{\pi}(x)$. The figure also reports the AUC, or area under the ROC curve, a common quantitative measure of a classifier's quality. The purple and orange points illustrate the quality of PredictFix alerts as predictors of breakdowns. For each point, the x-coordinate indicates the predictor's false positive rate (the proportion of non-breakdown weeks that have a PredictFix alert), and the x-coordinate indicates the predictor's true positive rate (the proportion of breakdown weeks that have a PredictFix alert). The purple point corresponds to high-severity alerts, and the orange point corresponds to medium-severity alerts. For each, the FPR and TPR were calculated using the observed breakdown outcomes in the data. The ROC curve, which illustrates the quality of our estimated π as a predictor of breakdowns, is included for comparison, as this represents the upper bound on the quality of any binary predictor that can be created (based on our data). All measures are constructed using data from the post period (as there are no PredictFix alerts in the pre period).

alert. Looking at the x-coordinates (false positive rate), we can see that, in contrast, only 2.3% of non-breakdown truck-weeks have a high-severity alert. Moreover, this point lies quite close to (though not exactly on) the ROC curve. In short, a high-severity PredictFix alert is a good indicator of breakdown risk.

Perhaps more surprisingly, the point corresponding to medium-severity PredictFix alerts lies much farther away from the ROC curve. While its location far above the 45-degree line indicates that it is still informative about breakdown risk, the clear contrast between the quality of the high- and medium-severity PredictFix alerts is a surprising feature of the setting. Although studying the design of the PredictFix algorithm is beyond the scope of this paper, Appendix B.2 includes a discussion of possible reasons for this non-optimality.

These results show that PredictFix provides technicians with one nearly-optimal binary predictor of breakdown risk. Is a single optimal binary predictor all the technician needs, or would she benefit from combining this predictor with other variables related to breakdown risk (e.g., sensor and fault variables)? The model presented in Section 5 helps to clarify this. Using this model, we show in Appendix B.3 that—in most cases—other predictors still play an important role.

4.3 Fact 3: Technicians respond to PredictFix but also ignore many alerts.

Having established that high-priority PredictFix alerts are a high-quality predictor of breakdown risk, we now ask how technicians respond to these alerts. To do so, we estimate an event study regression that captures the technician's propensity to do an engine repair around the time of an alert:

$$\text{Repair}_{i,t} = \alpha_0 + \sum_{k \in \{\text{pre},\text{post}\}} \sum_{\tau=-3}^{5} \beta_\tau^k \mathbb{1}\{t \in \mathcal{T}_k\} \widehat{\text{PredictFix}}^{\text{high}}_{i,t-\tau} + \alpha_i + \gamma_t + \epsilon_{i,t} \tag{1}$$

where $\text{Repair}_{i,t}$ is an indicator for a technician doing an engine repair on truck i in week t; $\mathcal{T}_{\text{pre}}$ and $\mathcal{T}_{\text{post}}$ represent the sets of weeks in the pre period and post period, respectively; and α_i and γ_t represent truck and week fixed effects, respectively. The variables of interest on the right-hand side are leads and lags of indicators for high-severity PredictFix alerts.[17] Note that we seek to estimate the responses to these PredictFix alerts (captured by the coefficients $\{\beta_\tau^k\}$) *separately* for the pre and post periods.

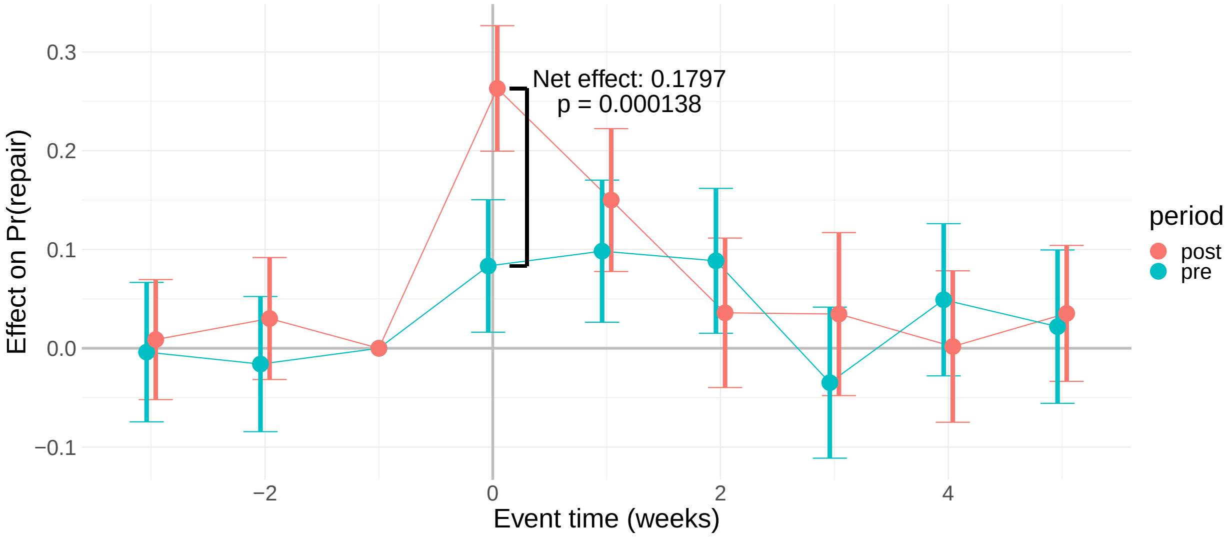
Because we do not observe actual PredictFix alerts in the pre period, we instead use a predictor of PredictFix alerts in estimating (1).[18] To form this predictor, we train a GBDT model

[17] We present the results of the corresponding analysis for medium-severity PredictFix alerts in Appendix B.4.1.

[18] For the analysis of the post period using actual, rather than predicted, PredictFix alerts, see Appendix B.4.2.

to predict high-severity PredictFix alerts using the x variables. The trained model predicts PredictFix alerts in out-of-sample data with a very high degree of accuracy; the ROC curve illustrating the quality of this predictor is presented in Figure 13 in Appendix D.4.[19]

Figure 4: Event Study: Response of repairs to (predicted) PredictFix alerts



Notes: This figure shows the estimated coefficients $\{\beta_\tau^k\}$ from equation (1). β_{-1}^{pre} and β_{-1}^{post} are normalized to zero, although the inclusion of week fixed effects means that differences in the average probability of repair across the pre and post periods are absorbed. In estimating (1), we use only observations not used in training the GDBT model that predicts PredictFix alerts. This means that we use all observations from the pre period, but only test sample observations from the post period. This eliminates the asymmetry between the in-sample/out-of-sample composition of the pre-period and post-period data that would result if we estimated (1) on the full sample.

Figure 4 presents the estimated event study coefficients $\{\beta_\tau^k\}$ from (1).[20] In this figure, we see that $\hat{\beta}_0^{\text{pre}} = 0.083$, which indicates that, in a pre-period week when a PredictFix alert *would have* occurred, the technician is 8.3pp more likely to do an engine repair than she was the week before. Since there were no PredictFix alerts in the pre period, this is not evidence of technicians responding to alerts; rather, it indicates that technicians in the pre period responded

[19] The predictor generated by the trained model has out-of-sample AUC-ROC measures of 0.977.

[20] While the outputs of the GBDT models are continuous predictors of the probability of a PredictFix alert, this event study exercise calls for binary predictors, i.e., indicators for whether a PredictFix alert would have occurred for truck i in week t. Converting a continuous predictor to a binary predictor requires selecting a threshold $\eta \in [0,1]$. We select the value of η that makes the predicted alert frequency equal to the actual frequency of alerts in the post period. A valid concern with this regression might be that noise in the predicted PredictFix variable (resulting from the fact that our prediction of alerts is imperfect) might bias the event study estimates. In general, we would expect such noise to bias the estimated effects toward zero. However, comparing the estimated effects in Figure 4 to those in Figure 10 (which is estimated using *actual* PredictFix alerts), we see that the estimated post-period same-week effect is actually slightly larger in the former. The fact that using $\widehat{\text{PredictFix}}_{i,t}$, rather than $\text{PredictFix}_{i,t}$, as a regressor results in a larger, rather than smaller, estimate of β_0^{post} speaks against the notion that prediction error plays a significant role in shaping our event study estimates.

to patterns in other signals (e.g., sensor measurements and fault codes) that PredictFix would identify as high risk. In the post period, we see that $\hat{\beta}_0^{\text{post}} = 0.263$, significantly larger than the pre-period response; the difference of 18.0pp indicates that PredictFix alerts change technician behavior, with technicians responding to alerts with an increased propensity to perform engine repairs.

In relative terms, the estimated 18.0pp response is large, as the unconditional probability of a repair for a given truck-week is 10.5%. However, in absolute terms, this response is small. It is certainly *not* the case that a PredictFix alert *always* leads to a repair; in fact, the probability of a repair for truck i in week t conditional on a high-severity PredictFix alert for truck i in week t is only 36.6%.

There are multiple possible interpretations of this observation. First, this apparently low degree of responsiveness may be the result of technicians exhibiting *discretion*. It may be optimal for the technician to consider other factors (e.g., costs) in combination with the PredictFix alert to determine if the risk of breakdown justifies a repair. Then, in some cases, the technician may opt not to do a repair when she sees a PredictFix alert because, for instance, the cost of doing a repair is very high. Second, this observation is also consistent with various behavioral explanations. For instance, the technician could be inattentive, always responding to a PredictFix alert when she notices one, but failing to notice most alerts. Alternatively, the low degree of responsiveness could be the result of the technician misunderstanding the process that generates PredictFix alerts.

These two kinds of explanations —-optimal use of discretion and mistakes resulting from behavioral forces —-are both plausible ex ante, but have vastly different normative implications. In developing and estimating the structural model of technician decision-making in Sections 5 and 6, our primary goal will be to separately estimate costs and technician beliefs about breakdown risk. Doing so will allow us to untangle these two explanations, drawing conclusions about the quality of technician decision-making and the extent to which it is improved by PredictFix.

4.4 Fact 4: No change in aggregate outcomes.

Having established that technicians are—at least somewhat—responsive to PredictFix alerts, we examine whether any aggregate effects of PredictFix—either positive or negative—are immediately apparent in the data. Using a simple linear probability model, we compare the average frequency of repairs and breakdowns in the pre and post periods:

$$y_{it} = \beta_0 + \beta_{\text{Post}} \mathbb{1}\{t \in \mathcal{T}_{\text{post}}\} + \epsilon_{it}, \tag{2}$$

where y_{it} is an indicator for an event—either a repair or a breakdown—for truck i in week t. The coefficient of interest, β_{Post}, captures the difference in pre-period versus post-period repair/breakdown frequencies. Estimates, which are presented in Table 3, show no statistically significant difference in repair or breakdown frequency across the periods.

Table 3: Differences in frequency and responsiveness across periods

	(1)	(2)
	Repair	Breakdown
Post	-0.00270	0.00305
	(0.00496)	(0.00206)
Constant	0.108***	0.0143***
	(0.00436)	(0.00181)
N	22127	22127

Notes: Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The two columns of this table report estimates for (2) with $y_{it} = \text{Repair}_{it}$ and $y_{it} = \text{Breakdown}_{it}$, respectively.

In light of our previous facts, this finding is perhaps surprising. Fact 3 indicates that technicians respond —- at least to some extent —- to PredictFix alerts. Fact 2 indicates that these alerts are high-quality predictors of breakdown risk. Taken together, these findings would seem to suggest that technicians in the post period would be more responsive to breakdown risk and that this might lead to better outcomes (fewer repairs and/or fewer breakdowns). However, Table 3 indicates that this is not the case; if anything, the frequency of breakdowns seems to be slightly higher in the post period. Fact 5, which offers evidence on differences in costs between the pre and post periods, helps rationalize these findings.

4.5 Fact 5: Repair costs are higher in the post period.

Regression (2) in the previous subsection provides a simple comparison of pre-period and post-period frequencies. This means that, rather than just capturing the effects of PredictFix, the estimated coefficient β_{Post} also captures the effects of other factors that changed between the pre and post periods. Given the manifold effects of the COVID-19 pandemic–on demand patterns, labor availability, and truck component availability–it seems plausible that technicians in the post period faced higher and more volatile costs.

To investigate this possibility, we once again estimate regressions of the form of (2), regressing two cost-related variables on a post-period indicator. Table 4 presents the estimated coefficients from these two regressions.

Table 4: Differences in repair costs between the pre and post periods

	(1) Tangible cost ($)	(2) Open work orders per truck
Post	128.1***	0.108***
	(34.75)	(0.00454)
Constant	599.3***	0.739***
	(30.21)	(0.00403)
N	8089	21331

Notes: Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The two columns of this table report estimates for (2) with y_{it} = Tangible cost of work order$_{it}$ and y_{it} = Number of open work orders$_{it}$/Number of trucks$_{it}$, respectively. The first column speaks to pre-post differences in tangible costs, as captured by the costs of work orders listed in the fleet maintenance records. The second column speaks to pre-post differences in the bindingness of capacity constraints, whose shadow costs represent a component of intangible costs. Both regressions include facility fixed effects.

First, to analyze changes in tangible repair costs, such as labor and materials, we regress work order costs from the maintenance records on a post-period indicator and facility fixed effects. The results, presented in column (1) of Table 4, show that tangible repair costs were, on average, $128 (20.6%) higher the post period.[21]

Second, for the intangible component of repair costs, we consider facility busyness as measured by the ratio of a facility's open work orders to the number of trucks assigned to that facility. If the repair capacity at a facility is proportional to its number of assigned trucks, this ratio indicates the utilization of repair capacity. A higher ratio suggests that the facility's capacity constraint is closer to binding, imposing a shadow cost of repair. The results, presented in column (2) of Table 4, show that this measure is higher on average in the post period.

Both findings indicate that technicians faced higher repair costs in the post period than in the pre period. Given our objective of evaluating PredictFix's effect on technician decision-making, these findings highlight the importance of controlling for inter-period cost differences. The structural model in the next section allows us to do this, quantifying the effects of introducing PredictFix while keeping cost conditions constant.

5 Model of technician decision-making

In this section, we develop a structural model of technician decision-making centered around a trade-off between the cost of doing a repair and the technician's perceived risk of breakdown.

[21] Notably, cost heterogeneity is also greater in the post period: the interquartile range for work order costs is $585, compared to $471 in the pre period.

We begin by presenting a simple static version of the model and describing the assumptions required for identification. Next, we introduce dynamics to account for the effect of a repair decision on the state of the truck in future periods. We describe the assumptions needed to feasibly estimate this dynamic model using high-dimensional data.

5.1 Static model

In a given week, a technician faces a binary choice, $a \in \{0,1\}$: she must decide whether to repair the truck ($a = 1$) or not ($a = 0$). Her payoff $u(a,s;v)$ depends on her repair decision a as well as the unknown potential breakdown outcome $s \in \{0,1\}$ and cost-related variables v. Without loss of generality, we normalize the payoff from no breakdown and no repair to zero: $u(0,0;v) = 0$ for all v. If the technician chooses not to do a repair and there is a breakdown, she incurs breakdown cost B: $u(0,1;v) = -B$.[22] If, on the other hand, the technician chooses to do a repair, any potential breakdown is prevented. In this case, payoffs do not vary with s, and she simply pays the repair cost: $u(1,0;v) = u(1,1;v) = -c(v)$. The payoffs for $(a,s) \in \{0,1\}^2$ are presented in Table 5.

Table 5: Static payoffs

		Potential breakdown	
		$s = 0$	$s = 1$
Repair action	$a = 0$	0	$-B$
	$a = 1$	$-c(v)$	$-c(v)$

Since payoffs depend on the potential breakdown outcome s, the technician's decisions depend critically on her beliefs about the likelihood of a breakdown. The technician observes x, a high-dimensional variable describing the state of the truck, and uses x to form a prediction about the probability of breakdown. We *do not* impose a specific model of how the technician forms her predicted probability of breakdown, which we denote $\boldsymbol{\rho}(x)$; we allow $\boldsymbol{\rho}(x)$ to differ arbitrarily from the true probability of breakdown, which we denote $\boldsymbol{\pi}(x)$.

Multiple forces could generate this divergence. First, the technician may not fully observe x, perhaps due to limited attention and the complexity and high-dimensionality of x. Second, even if the technician does fully observe x, she may make mistakes in mapping x to breakdown probability. For instance, the technician could incur some cost of effort for computation

[22] The assumption that the breakdown cost is constant reflects thinking of the team responsible for PFC's fleet maintenance operations. When asked how both they and the technicians perceive variation in breakdown costs, they stated that neither they nor the technicians take this variation into account when making decisions. Note that we are *not* asserting that we know the exact value of B or that the technician's perceived cost of breakdown B is the true cost of a breakdown.

and optimally choose to exert a level of effort less than that necessary to have $\boldsymbol{\rho}(x) = \boldsymbol{\pi}(x)$. Alternatively, various behavioral biases or lack of experience could prevent the technician from understanding the process that generates breakdowns. We take no position on which of these forces shape the technician's predictions and for now allow $\boldsymbol{\rho}$ to differ arbitrarily from $\boldsymbol{\pi}$. Note that while we are agnostic as to whether the technician's breakdown risk predictions $\boldsymbol{\rho}$ are correct, we assume that she knows the costs $c(v)$ and B.[23]

Having formed a prediction $\boldsymbol{\rho}(x)$, the technician then performs a repair if and only if

$$\boldsymbol{\rho}(x) > \frac{c(v)}{B} \equiv \tau(v) \tag{3}$$

The function $\tau(v)$ captures the technician's breakdown-risk threshold and is equal to the ratio of the cost of a repair to the cost of a breakdown.

The simple decision rule in equation (3) shows that the technician's choices are determined by just two objects: the cost ratio $\tau(v)$ and the technician's breakdown-risk prediction $\boldsymbol{\rho}(x)$. To use this model as a tool to understand the data and how PredictFix changes decision-making, we would like to estimate both of these objects. However, here we encounter a common econometric challenge: separately identifying preferences and beliefs. Overcoming this challenge will require three steps: an exclusion restriction, a restriction on private information, and a restriction to identify level and scale.

First, we impose an exclusion restriction. The intuition for identifying the model from the choice data is that $\boldsymbol{\rho}$ and τ will be identified by observing how repair choices respond to variation in x and variation in v, respectively. If there were some variable z that were in both x and v and we observed the response of repair decisions to variation in z, we would not know whether to attribute that response to the effect of z' on beliefs or the effect of z' on costs. Assumption 1 states the condition required to overcome this challenge:

Assumption 1 (Exclusion restriction)**.** *The set of variables in v and the set of variables in x are mutually exclusive.*

This assumption states that there are no variables observed by the technician that affect both costs and breakdown-risk predictions. In our context, this is a natural assumption, as the variables in w are purely economic in nature; w includes both shifters of *intangible* costs (e.g., a facility's capacity for doing repairs) and $\hat{c}^{\text{tangible}}$, a measure of *tangible* costs constructed from observed costs recorded in the maintenance records.[24] On the other hand, x—which includes

[23] This asymmetry between the technician's knowledge of breakdown risk (potentially imperfect) and her knowledge of costs (perfect) is driven by our question of interest: We want to understand how technician decision-making is changed by PredictFix, a tool designed to predict breakdowns, not costs.

[24] For a description of how $\hat{c}^{\text{tangible}}$ is constructed, see Appendix D.

sensor measurements and fault codes describing the physical state of the truck's components, its usage, and its environmental conditions—contains all of the variables that might reasonably affect the truck's risk of breakdown.[25] For a full list of the variables in x, see Table 1 in Section 3.2; for a full list of the variables in w, see Table 9 in Appendix D.

Second, we impose a restriction on private information. While we can identify a model in which the technician has private information on either costs or beliefs, we cannot identify a model in which the technician has private information on both. We therefore face a choice of which of the two to impose a restriction on. In many settings, both options would be unattractive; however, the features of our setting make it natural to restrict the information underlying beliefs. Recall that one of the extraordinary features of our setting is that we observe a rich set of sensor measurement and fault code data that describe the performance of various truck components, the way the truck is being used, and the environmental conditions in which it is operating. Because, taken together, these data provide us with a comprehensive description of the state of the truck, it is natural to assume that no private information enters the technician's beliefs about breakdown risk. Formally,

Assumption 2 (No private information)**.** *Suppose ξ is observed by the technician but not by the econometrician. Then,* $\Pr(\textit{breakdown} \mid x, \xi) = \Pr(\textit{breakdown} \mid x) = \boldsymbol{\pi}(x)$.

However, we can still allow for private information on costs. In particular, we assume that this private information takes the form of an additive idiosyncratic cost shock:

$$\tau(v) = g(w) + \epsilon \tag{4}$$

where $v = (w, \epsilon)$, $\epsilon \sim$ iid Logistic(θ), and w is observable by the econometrician. This implies that the probability of a repair given observables (w, x) is

$$p(w, x) \equiv \Pr(a = 1 \mid w, x) = \Lambda(\theta[-g(w) + \boldsymbol{\rho}(x)]) \tag{5}$$

where Λ is the Logistic function.

Finally, to identify both levels and scales of costs (g) and beliefs ($\boldsymbol{\rho}$), we require two restrictions on beliefs.[26] To see why, note two features of equation (5). First, $-g(w)$ and $\boldsymbol{\rho}(x)$ are added together, which means that g and $\boldsymbol{\rho}$ are each only identified up to an additive constant. Second, the scale parameter θ multiplies both $g(w)$ and $\boldsymbol{\rho}(x)$, meaning that θ, g and $\boldsymbol{\rho}$ are each only identified up to a multiplicative constant. We can achieve separate identification by imposing two mild restrictions on beliefs.

[25] Any effect of these variables on tangible costs should be effectively controlled for by $\hat{c}^{\text{tangible}}$.

[26] Both costs and beliefs are partially identified under Assumptions 1 and 2 alone. For the formal result, see Appendix C.1.

Assumption 3 (Beliefs mean and minimum)**.** *The technician's breakdown risk prediction function $\boldsymbol{\rho}$ satisfies the following conditions:*

- *The technician's beliefs are correct on average, i.e.,* $\mathbb{E}_x\boldsymbol{\rho}(x) = \mathbb{E}_x\boldsymbol{\pi}(x)$*.*
- *There is some $x \in \mathcal{X}$ for which the technician believes that the risk of breakdown is zero, i.e.,* $\min_x \boldsymbol{\rho}(x) = 0$*.*

The first of these restrictions says that, while technician's perceived risk of breakdown for any given state x may be wildly different from the true risk of breakdown, the average perceived risk of breakdown over all states is equal to the average true risk of breakdown. The second restriction imposes that there is some state x for which the technician believes there is no risk of breakdown. Although these two restrictions put limits on the technician's perceived breakdown risk, $\boldsymbol{\rho}$ remains largely unrestricted. In particular, Assumption 3 does not impose any restrictions on how the technician *orders* states by risk. This flexibility is particularly important, as this ordering is what will determine the results of our ROC curve-based analysis of how PredictFix changes the technician's perceived risk of breakdown (see Section 7.1).

Under this set of assumptions, the static model is identified from choice data:

Proposition 1 (Static identification)**.** *Consider the static choice model. Suppose $g(\cdot)$ is a differentiable function whose domain is a compact connected set $\mathcal{W}$, and Assumptions 1-3 are satisfied. If the support of w conditional on each $x \in \mathcal{X}$ is equal to $\mathcal{W}$, then θ, $g(\cdot)$ and $\boldsymbol{\rho}(\cdot)$ are identified.*

Proof. See Appendix C.2.2. □

5.2 Dynamic Model

While the static model in the previous subsection captures the technician's key trade-off—between repair cost and breakdown risk—it does not account for the effect of the current repair decision on the state in future periods. In this subsection, we rectify that shortcoming by incorporating dynamics into the model. We describe a set of assumptions under which (a) the model is identified from dynamic choice data and (b) the model can be estimated from such data, even when the state variable is high-dimensional.

Accounting for the technician's dynamic considerations requires taking a stand on the technician's beliefs about the future. Recall that in the static model, the technician's beliefs about the risk of breakdown may be incorrect, i.e., $\boldsymbol{\rho}(x) \overset{?}{=} \boldsymbol{\pi}(x)$. We naturally assume that technicians also use their (potentially incorrect) beliefs about breakdown risk when evaluating potential future payoffs; for example, if x_{it+1} is a possible future state of truck i at time $t+1$, the technician at time t believes that the probability of breakdown in that state is $\boldsymbol{\rho}(x_{it+1})$.

While the technician's predictions of breakdown risk may be flawed, her understanding of the state transition process is not. That is, she knows the true distribution of the next period's state (w_{it+1}, x_{it+1}) conditional on the current state (w_{it}, x_{it}) and her action a_{it}. This assumption is not only standard, but necessary. The model could be identified with *either* technicians having incorrect beliefs about breakdown risk *or* technicians having incorrect beliefs about the transition process, but not both. Given that this paper focuses on how a predictive algorithm shapes technicians' predictions and decisions, assuming that technicians know the true transition process is the natural choice.

Introducing dynamics adds one additional term to the expression for the conditional probability of repair, reflecting the perceived effect of a repair in period t on future discounted payoffs.

$$\begin{aligned} p(w_{it}, x_{it}) &= \Lambda\left(\theta\left[v_1(w_{it}, x_{it}) - v_0(w_{it}, x_{it})\right]\right) \\ &= \Lambda\left(\theta\left[-g(w_{it}) + \boldsymbol{\rho}(x_{it}) + \delta\left(EV_1(w_{it}, x_{it}) - EV_0(w_{it}, x_{it})\right)\right]\right) \end{aligned} \tag{6}$$

where, following standard notation from the literature on dynamic models with Markov transition processes, $v_a(w_{it}, x_{it})$ represents the technician's payoff from action a, inclusive of the continuation value (discounted by δ) but excluding the idiosyncratic cost shock ϵ; $EV_a(w_{it}, x_{it})$ represents the ex-ante expected value function following action a, i.e., if the state in period t is (w_{it}, x_{it}) and the technician chooses action a, her expected continuation value (before w_{it+1}, x_{it+1}, and ϵ_{it+1} are realized) is $EV_a(w_{it}, x_{it})$. Therefore, the difference $EV_1(w_{it}, x_{it}) - EV_0(w_{it}, x_{it})$, captures the technician's perceived effect of the current action on expected future payoffs.

Keeping in mind our goal of estimating beliefs $\boldsymbol{\rho}$ and costs τ, the question is how to take equation (6) to the data. As is typically the case with dynamic discrete choice models, the ex-ante value functions EV_0, EV_1 are not directly observable and must be computed. However, the fact that our state variable (w_{it}, x_{it}) is high-dimensional rules out some standard approaches to dealing with these ex-ante value functions. In particular, the nested fixed-point approach of Rust (1987) is infeasible.

Instead, we will proceed by leveraging the insights of Hotz and Miller (1993), who showed that the ex-ante value functions can be expressed in terms of choice probabilities. Using our logistic-distributed cost shock assumption, the difference of ex-ante value functions can be written as a sum of three terms:

$$
\begin{aligned}
EV_1(w_{it}, x_{it}) - EV_0(w_{it}, x_{it}) = {} & \mathbb{E}[v_1(w_{it+1}, x_{it+1}) \mid w_{it}, x_{it}, a_{it} = 1] \\
& - \mathbb{E}[v_1(w_{it+1}, x_{it+1}) \mid w_{it}, x_{it}, a_{it} = 0] \\
& - \frac{1}{\theta}\mathbb{E}[\log(p(w_{it+1}, x_{it+1})) \mid w_{it}, x_{it}, a_{it} = 1] \\
& + \frac{1}{\theta}\mathbb{E}[\log(p(w_{it+1}, x_{it+1})) \mid w_{it}, x_{it}, a_{it} = 0] \\
= {} & \underbrace{(\mathbb{E}[g(w_{it+1}) \mid w_{it}, x_{it}, a_{it} = 0] - \mathbb{E}[g(w_{it+1}) \mid w_{it}, x_{it}, a_{it} = 1])}_{\Delta Eg(w_{it}, x_{it})} \\
& + \underbrace{\big(\mathbb{E}[EV_1(w_{it+1}, x_{it+1}) \mid w_{it}, x_{it}, a_{it} = 1] - \mathbb{E}[EV_1(w_{it+1}, x_{it+1}) \mid w_{it}, x_{it}, a_{it} = 0]\big)}_{\Delta EV(w_{it}, x_{it})} \\
& + \frac{1}{\theta}\underbrace{(\mathbb{E}[\log(p(w_{it+1}, x_{it+1})) \mid w_{it}, x_{it}, a_{it} = 0] - \mathbb{E}[\log(p(w_{it+1}, x_{it+1})) \mid w_{it}, x_{it}, a_{it} = 1])}_{\Delta E \log p(w_{it}, x_{it})}
\end{aligned}
\tag{7}
$$

The first equality is an application of the Hotz and Miller (1993) inversion. The second equality uses the fact that $v_1(w, x) = -g(w) + \Delta EV_1(w, x)$. The three terms in equation (7) capture the expected effect of a repair on, respectively, the next period's cost ratio, the next period's ex-ante expected value function for $a = 1$, and the log probability of repair in the next period. Thinking about estimation, the first and third terms are manageable, but the second one is not. Yet, as described by Arcidiacono and Miller (2011) and Arcidiacono and Ellickson (2011), this term is equal to zero in settings with *finite dependence*. Our setting can be thought of as one that has this feature if a repair is a *renewal action* that resets all of the non-exogenously-evolving components of the state variable. Assumption 4 formalizes this idea:

Assumption 4 (Repair is a renewal action)**.** *The transition process of each component z of the state variable (w, x) satisfies* either *of the following conditions:*

(i) z resets after a repair, i.e., the conditional distribution of $z_{it+1} \mid a_{it} = 1, w_{it}, x_{it}$ does not depend on (w_{it}, x_{it}); or

(ii) z evolves independently and exogenously, i.e., the conditional distribution of $z_{it+1} \mid a_{it}, w_{it}, x_{it}$ only depends on z_{it}.

Moreover, at least one element of x satisfies (i), and all elements of w satisfy (ii).

For each component of our state variable (w, x), either condition (i) or (ii) is plausible. First, all components of w—which includes the facility's number of open work orders and the facility's number of trucks, as well as month and facility fixed effects—plausibly satisfy (ii). Second and similarly, in x, weather-related sensor measurements (e.g., outside air temperature and pressure) satisfy (ii) as do, (at least to a first approximation) truck odometer mileage.[27] Finally, the remaining components of x—which are primarily fault and sensor variables describing the status of the engine—are likely to reset after a repair (and therefore satisfy (i)), as a repair should (ideally) address any active issues.[28]

The benefits of Assumption 4 are twofold. First, the main benefit is that Assumption 4 implies that at time t, the expected distribution of (w_{t+2}, x_{t+2}) is the same whether the technician chooses $(a_{it}, a_{it+1}) = (0,1)$ or $(a_{it}, a_{it+1}) = (1,1)$. As a consequence, the second term on the right-hand side of equation (7) is zero. Second, the assumption that all components of w evolve exogenously means that a repair in the current period has no effect on the cost ratio in future periods. Therefore, the first two terms in equation (7) drop out, and the difference in ex-ante expected values becomes

$$EV_1(w_{it}, x_{it}) - EV_0(w_{it}, x_{it}) = \frac{1}{\theta} \Delta E \log p(w_{it}, x_{it}) \tag{8}$$

That is, the difference in the ex-ante value functions can be written as $1/\theta$ times the (negative of the) expected effect of a repair this period on next period's log repair probability. The conditional repair probability is then

$$p(w_{it}, x_{it}) = \Lambda\left(-\theta g(w_{it}) + \theta \rho(x_{it}) + \delta\left[\theta \Delta E g(w_{it}, x_{it}) + \Delta E \log p(w_{it}, x_{it})\right]\right) \tag{9}$$

Let us now consider what bringing equation (9) to the data would require. Note that computing $\Delta E \log p(w, x)$ requires integrating the conditional choice probability function over the transition process for (w, x). With a high-dimensional x, it is not feasible to flexibly estimate—and, therefore, integrate over—the state transition process. However, we show that under Assumption 5, $\Delta E p(w, x)$ can be expressed exactly in terms of choice probabilities that are more easily estimated from the data.

Assumption 5. *The transition process for state variables (w_{it}, x_{it}) and the technician's conditional choice probability function $p(\cdot,\cdot)$ are such that*

$$p_{it+1}(w_{it+1}, x_{it+1}) \mid a_{it}, w_{it}, x_{it} \sim \text{Beta}\left(\mu(a_{it}, w_{it}, x_{it}), \nu\right)$$

where $\mu : \{0,1\} \times \mathcal{W} \times \mathcal{X} \to \mathbb{R}^+$ and $\nu \in \mathbb{R}^+$.[29]

[27] Of course, if a technician takes a truck off the road to do a repair, that will affect the odometer mileage. However, technicians' repair decisions likely only respond to large differences in mileage (e.g., 250,000 miles versus 200,000) rather than the relatively small differences that might result from a truck being off the road a few days while it is repaired.

[28] x includes the number of weeks and the number of miles since last repair, both of which, by definition, reset when a repair is done.

[29] Here we use "mean-precision" parameterization of the Beta distribution, as this is more convenient for our

Assumption 5 imposes a joint restriction on *both* the transition process and the choice probability function p. Such assumptions may, in some contexts, be inadvisable, as it can be difficult to evaluate their plausibility. However, the Beta distribution is relatively flexible. Any plausible continuous conditional distribution of p_{it+1} is likely well-approximated by a Beta distribution with some parameter values. In addition, Assumption 5 allows one of the two parameters of the distribution to vary arbitrarily with the current state and action.

The benefit of Assumption 5—that it allows $\Delta E \log p(w_{it}, x_{it})$ to be estimated without estimating the transition process—is derived from the following properties of the Beta distribution: First, if $p \sim \text{Beta}(\mu, \nu)$, then the mean of the distribution is $\mathbb{E}[p] = \mu$. Second, the expectation of $\log p$ is $\mathbb{E}[\log p] = \psi(\mu\nu) - \psi(\nu)$, where ψ is the digamma function. This second property implies that, under Assumption 5, the expected log of the next period's choice probability has a closed-form expression that is a function of the distribution's two parameters. Since the dynamic term $\Delta E \log p(w_{it}, x_{it})$ is a difference of two such expected logs, this means that the dynamic term can be expressed in closed form, rather than as an integral over a high-dimensional state variable:

$$\Delta E \log p(w_{it}, x_{it}) = \psi(\mu(0, w_{it}, x_{it})\nu) - \psi(\mu(1, w_{it}, x_{it})\nu)$$

Apart from Beta (and the closely related Kumaraswamy distribution), we know of no other distributional assumption that allows for $\Delta E \log p(w_{it}, x_{it})$ to be written in closed form.[30]

The fact that μ parameterizes the distribution's mean indicates that $\mu(a_{it}, w_{it}, x_{it})$ is equal to the expectation of the next period's choice probability, which is identified from the data. Moreover, this choice probability—despite the high-dimensionality of the state—is much more credibly estimable than the transition process.

Proposition 2 shows that the second distributional parameter, ν, along with the static payoff primitives—θ, γ, and $\boldsymbol{\rho}$—is identified from the technician's dynamic repair choices.

Proposition 2. *Consider the dynamic choice model. Suppose $g(\cdot)$ is a differentiable function whose domain is a compact connected set $\mathcal{W}$, and Assumptions 1-5 and Technical Condition 1 are satisfied. If the support of w conditional on each $x \in \mathcal{X}$ is equal to $\mathcal{W}$, then θ, ν, $\mu(\cdot,\cdot)$, $g(\cdot)$ and $\boldsymbol{\rho}(\cdot)$ are identified.*

Proof. See Appendix C.2.3. □

purpose than the more standard (α, β) parameterization. To convert from the former to the latter, $\alpha = \mu\nu$ and $\beta = (1-\mu)\nu$.

[30] There are of course, a few trivial examples (e.g., $p_{it+1} \sim \text{Uniform}[a, b]$) that would also have this property, but would impose much more extreme (and probably unrealistic) assumptions on the conditional distribution of p_{it+1}.

Technical Condition 1, stated in Appendix C.2.3, imposes a mild technical condition necessary for proving identification of the parameter ν. It is easily satisfied, as it need only hold for *some* state (x, w) in the vast high-dimensional state space $\mathcal{X} \times \mathcal{W}$.

Before proceeding to the details of estimation and presenting our results, it is useful to pause and take stock of the assumptions we have and have not made up to this point. This discussion serves to clarify the conditions under which our estimates are or are not valid.

The most substantive restrictions that we have imposed are as follows: *First*, in Assumption 2, we asserted that there is no private, breakdown-relevant information observable to the technician but not to us. *Second*, in Assumption 3, we asserted that the technician knows the mean breakdown risk and believes the minimum breakdown risk to be zero. *Third,* in Assumption 5, we asserted that the conditional distribution of the period-ahead choice probability is distributed according to the Beta distribution.

While the above assumptions are necessary for identification and/or estimation of the dynamic model, in other important respects, we have left technician behavior unrestricted. *First*, we have not asserted that the technician's objective is aligned with that of the firm. If agency problems exist or if the technician misunderstands the process that determines repair or breakdown costs, this poses no issue for us in terms of our identification and estimation of the model. *Second*, we have not imposed significant restrictions on technician preferences. In particular, whether technicians are risk-averse, risk-neutral, or risk-seeking has no bearing on our ability to identify or estimate the model.

If there is a misalignment of objectives or if technicians have non-risk-neutral preferences, this matters only insofar as it affects the *interpretation* of $\tau(\nu) = g(w) + \epsilon$. If the technician is risk neutral, then τ this has the interpretation of the ratio of the technician's perceived cost of repair to the technician's perceived cost of breakdown. If, furthermore, the technician's objective is aligned with that of the firm, then τ has the interpretation of the ratio of the true cost of repair to the true cost of breakdown.

6 Estimation

This section describes how we bring the dynamic model presented in the previous section to the data.

Before we can estimate the dynamic structural model, we must first carry out two pre-estimation steps. First, in recognition of the fact that our finite data limits the number of parameters that can be feasibly estimated, we perform a variable selection step, reducing the dimension of x from more than 2,000 to 20. For details, see Appendix D.1.2. Second, we estimate the conditional week-ahead choice probability function $\mu(a_t, w_t, x_t) = \Pr(a_{t+1} \mid a_t, w_t, x_t)$ offline using a GBDT model. For details, see Appendix D.1.3. For a full definition of the set of variables included in w and x, see Appendix D.1.1.

Recall that, in Proposition 2, we showed that the cost function g and the technician's perceived breakdown risk function $\boldsymbol{\rho}$ are nonparametrically identified from choice data. However, in bringing the model to the data, we introduce functional form assumptions that limit the number of parameters to be estimated. In particular, we impose that g is linear

$$g(w) = \gamma_0 + w'\gamma_1$$

and that $\boldsymbol{\rho}$ is logistic-linear

$$\boldsymbol{\rho}(x) = \Lambda\left(\lambda_0 + x'\lambda_1\right)$$

where Λ is the logistic function.

We estimate the model using a constrained maximum likelihood approach, estimating the model's parameters separately for the pre period and the post period. The constraints we impose correspond to the two restrictions on beliefs described in Assumption 3.[31] With these constraints, the optimization problem for each period is as follows:

$$\begin{aligned} &\max_{\theta,\nu,\gamma,\lambda} \sum_i \log \mathcal{L}\left(a_i, w_i, x_i \mid \theta, \nu, \gamma, \lambda\right) \qquad (10)\\ &\quad \text{subject to } \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\rho}(x_i;\lambda) = \bar{\boldsymbol{\pi}} \\ &\qquad\qquad\qquad \min_x \boldsymbol{\rho}(x;\lambda) \approx 0 \\ &\quad \text{where} \quad \mathcal{L}\left(a_i, w_i, x_i \mid \theta, \nu, \gamma, \lambda\right) = \begin{cases} 1 - p(w_i, x_i \mid \theta, \nu, \gamma, \lambda) & \text{if } a_i = 0 \\ p(w_i, x_i \mid \theta, \nu, \gamma, \lambda) & \text{if } a_i = 1 \end{cases} \end{aligned}$$

Solving this constrained maximization problem numerically is challenging because the objective function is not convex. This non-convexity arises from the fact that the choice probability expression (9) features one non-linear function, $\boldsymbol{\rho}$, inside of another non-linear function, the logistic function. With this non-convexity in mind, we choose a combination of optimization techniques that increase the likelihood of finding the global optimum. These include (1) a gradient-based momentum algorithm and (2) a multi-start method. To enforce the constraints, we use the augmented Lagrangian approach. For details of our approach to solving (10), see

[31] In practice, because of our chosen functional form, having $\boldsymbol{\rho}(x)$ exactly equal to zero is not possible, as the range of the logistic function is $(0,1)$. Therefore, the constraint that we actually impose is $\min_x \boldsymbol{\rho}(x) \leq 10^{-8}$.

Appendix D.3.

For estimation, as well as for the analysis of counterfactuals in Section 8, we assume $\delta = 0.999$. This is, admittedly, small for a *weekly* discount factor. However, this choice is dictated by computational constraints; as explained in Section 8, we analyze several hundred counterfactuals, each of which requires solving a Bellman equation fixed point problem for the ex-ante expected value functions EV_0, EV_1. The convergence of such problems becomes substantially slower as δ approaches 1.

7 Estimates: Breakdown-risk predictions and costs

By estimating the dynamic structural model, we recover two sets of key primitives—technicians' perceived breakdown risk ($\boldsymbol{\rho}$) and cost parameters (θ, γ)—for both the pre and post periods. In this section, we examine each of these sets of estimated primitives. *First*, in Section 7.1, we compare our estimates of technicians' perceived breakdown risk $\boldsymbol{\rho}$ without PredictFix (pre period) and with PredictFix (post period). Evaluating each as a predictor of actual breakdowns, we find that technicians with PredictFix exhibit a substantially better ability to predict breakdowns. *Second*, in Section 7.2, we compare the estimated costs for the pre and post periods. We find that the effective cost ratio in the post period has both a higher mean and a higher variance. This finding is consistent with Fact 5 from Section 4.5, which presented suggestive evidence of such pre-post cost differences and hypothesized that such differences could explain the seeming contradiction between Facts 2-3 and Fact 4.

For a discussion of model fit, see Appendix D.4.

7.1 The effects of PredictFix on technicians' predictions

We begin by analyzing our estimates of $\boldsymbol{\rho}$, the technician's beliefs about the probability of breakdown. To understand how PredictFix changes human prediction, we assess the quality of these beliefs, both with and without PredictFix ($\boldsymbol{\rho}_{\text{post}}$ and $\boldsymbol{\rho}_{\text{pre}}$, respectively), as predictors of actual breakdowns. We perform this analysis of beliefs in two ways: First, we again make use of a tool introduced in Section 4.1—ROC curves—to assess the extent to which PredictFix changes the technician's ability to *order* states by riskiness. Second, to understand how PredictFix changes the *calibration* of technicians' beliefs, we estimate a logistic regression of actual breakdown outcomes on the technician's perceived breakdown risk.

We begin by generating three different ROC curves, which are summarized in Table 6. The first two ROC curves speak to the quality of $\boldsymbol{\rho}_{\text{pre}}$ and $\boldsymbol{\rho}_{\text{post}}$, respectively, as predictors of

Table 6: ROC curve specifications

Predictor	Outcome	Sample restriction
$\boldsymbol{\rho}_{\text{pre}}$	Breakdowns	$a_{it} = 0$
$\boldsymbol{\rho}_{\text{post}}$	Breakdowns	$a_{it} = 0$
$\hat{\boldsymbol{\pi}}$	Breakdowns	Test sample and $a_{it} = 0$

breakdowns. In both, the sample is restricted to the set of truck-weeks without repairs.[32] While our primary interest is in the comparison between $\boldsymbol{\rho}_{\text{pre}}$ and $\boldsymbol{\rho}_{\text{post}}$, it is also useful to have, as a benchmark, the ROC curve for a high-quality predictor of breakdowns. To this end, the third ROC curve speaks to the quality of $\hat{\boldsymbol{\pi}}$, our GBDT-based predictor of breakdown risk described in Section 4.1, as a predictor of actual breakdowns.

These three ROC curves are presented together in Figure 5. We see that the red curve (corresponding to $\boldsymbol{\rho}_{\text{post}}$) is everywhere above the blue curve (corresponding to $\boldsymbol{\rho}_{\text{pre}}$). Recalling the PPF interpretation of the ROC curve discussed in Section 4.1, this means that $\boldsymbol{\rho}_{\text{post}}$ *strictly dominates* $\boldsymbol{\rho}_{\text{pre}}$ as a predictor of breakdowns.

Now comparing the quality of $\boldsymbol{\rho}_{\text{post}}$ with that of the benchmark predictor $\boldsymbol{\pi}$ (illustrated by the black curve), we see that, while technicians could still improve their prediction of breakdowns, technicians in the post period perform fairly well. Using the AUC as a quantitative measure of quality and considering the AUC for $\boldsymbol{\pi}$ as an upper bound on feasible predictive quality, we see that with the introduction of PredictFix, the gap between the prediction of breakdowns by technicians and the best possible prediction of breakdowns narrows by $\frac{0.704-0.598}{0.779-0.598} \approx$ 59%.

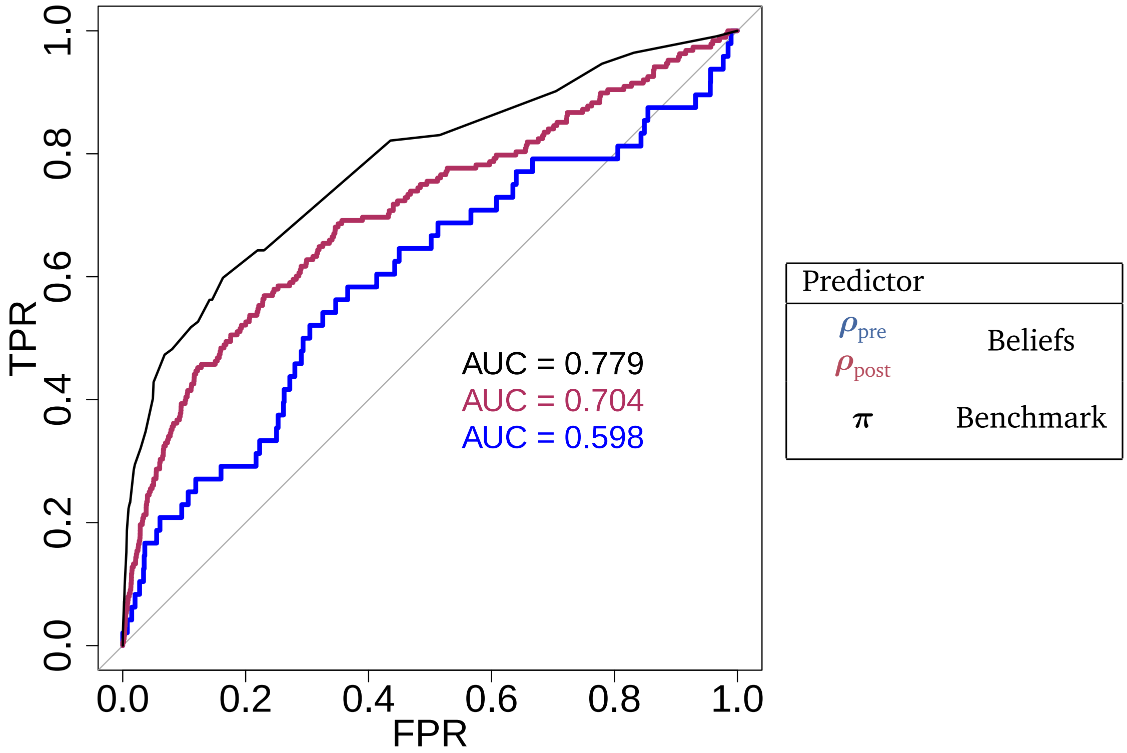
This relatively high quality of $\boldsymbol{\rho}_{\text{post}}$ as a predictor of breakdowns is notable not just because it represents a substantial improvement over the pre period, but also because it suggests that the model does a reasonably good job of capturing reality. The relationship between $\boldsymbol{\rho}$ and actual breakdown outcomes is an *untargeted moment* in the sense that no data on actual breakdowns was used in the estimation of the model.[33] For more on model fit, see Appendix D.4.

In interpreting these results, it is useful to note that an ROC curve only reflects a predictor's ability to *order* states correctly by risk; that is, the ROC curve is invariant to monotone increasing transformations of the predictor. On the one hand, this bolsters the robustness of the ROC results in Figure 5; the assumptions required to identify how the technician orders states by perceived breakdown risk are weaker than those required to fully identify $\boldsymbol{\rho}$. In particular,

[32] If a technician does a repair, this likely prevents any potential breakdown, so the potential breakdown outcome is unobserved when $a_{it} = 1$.

[33] The only exception is the very minimal contribution via the constraint that $\mathbb{E}\boldsymbol{\rho}(x) = \bar{\pi}$, which just affects the mean and should have little or no effect on the extent to which $\boldsymbol{\rho}$ correctly *orders* states in terms of breakdown risk.

Figure 5: ROC Curves: ρ_{pre} and ρ_{post} as predictors of breakdowns



Notes: This figure illustrates the quality of estimated beliefs ρ_{pre} and ρ_{post} as predictors of observed breakdowns. The ROC curves for pre and post beliefs are the blue and red curves, respectively. A nonparametric benchmark (out-of-sample fit for our GBDT estimate of π) is included in black.

Proposition C.1 in Appendix C.1 shows that, under only Assumptions 1-2, ρ is identified up to an affine, increasing transformation. This means that the main result of this section—that the ROC curve for ρ_{post} strictly dominates that for ρ_{pre}—is robust to relaxations of the restrictions on beliefs imposed by Assumption 3. However, the fact that the ROC curve only captures the extent to which a predictor correctly orders states by risk also means that Figure 5 does not capture the full picture of how PredictFix changes the quality of the technician's predictions; in particular, it does not reflect the extent to which ρ_{pre} and ρ_{post} are well-calibrated (i.e., correctly scaled) predictors of breakdown risk.

To analyze the calibration of beliefs, we estimate logistic regressions of actual breakdown outcomes on $\Lambda^{-1}(\rho)$, the technician's perceived breakdown risk ρ with the inverse-logistic transformation applied:

$$\Pr\left(\text{Breakdown}_{it} \mid \rho_j(x_{it})\right) = \Lambda\left(\phi_0 + \phi_1 \Lambda^{-1}\left(\rho_j(x_{it})\right)\right) \tag{11}$$

for $j = \text{pre, post}$. Applying the inverse-logistic transformation to beliefs makes the coefficients

Table 7: Logistic regression of actual breakdowns on technicians' perceived breakdown risk

	(1) Breakdown	(2) Breakdown
$\Lambda^{-1}\left(\rho_{\text{pre}}\right)$	0.233^{***}	
	(0.0435)	
$\Lambda^{-1}\left(\rho_{\text{post}}\right)$		0.814^{***}
		(0.0764)
Constant	-3.198^{***}	-0.817^{*}
	(0.218)	(0.318)
N	19091	19091

Notes: Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. This table presents coefficient estimates for the logistic regression described by equation (11). The two columns correspond to regressions with the predictors ρ_{pre}—the technician's beliefs without PredictFix—and ρ_{post}—the technician's beliefs with PredictFix, respectively. A predictor that captured the true risk of breakdown would have intercept zero and slope one. The estimates show that technician beliefs with PredictFix come much closer to this benchmark.

more interpretable; if the technician's beliefs exactly captured the true risk of breakdown (i.e., $\rho(x) = \pi(x)$ for all x), then we would have $\phi_0 = 0$ and $\phi_1 = 1$.

The estimated coefficients for this regression with ρ_{pre} and ρ_{post} are presented in Table 7. The results show that, while both ρ_{pre} and ρ_{post} are statistically significant predictors of breakdown risk, ρ_{post}—the technician's beliefs with PredictFix —- is a much better and better-calibrated predictor. This finding, combined with the results in Figure 5 above, represents strong evidence that PredictFix improves the ability of technicians to predict the risk of breakdown.

7.2 Cost ratio estimates

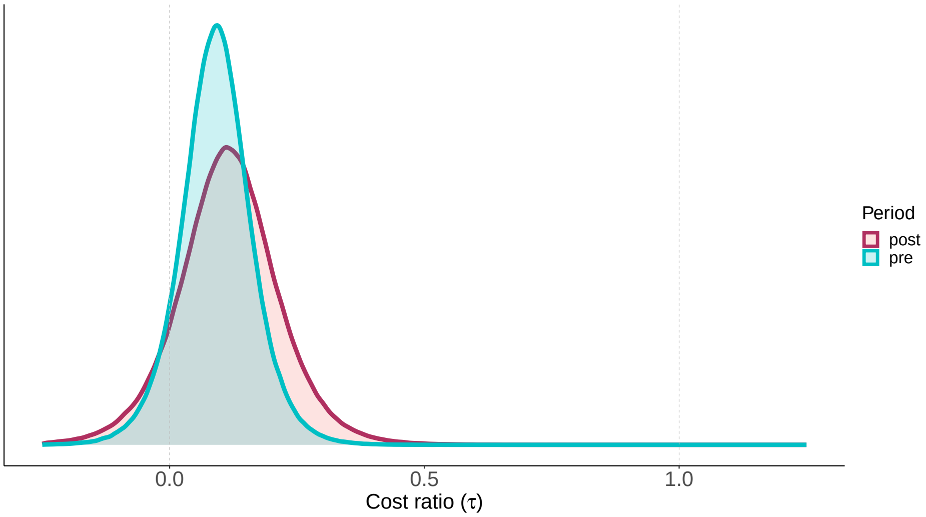
To better understand the differences in behavior and outcomes between the pre and post periods, we next explore our costs estimates for the two periods. Figure 6 shows the distribution of the cost threshold

$$\tau(v_{it}) = \gamma_0 + w_{it}'\gamma_1 + \delta\Delta E \log p(w_{it}, x_{it}; \nu) + \epsilon$$

constructed using the estimated cost parameters θ, γ and the estimated transition-process primitives μ, ν. The inclusion of the dynamic term reflects the fact that this term, like the static cost terms, shifts the technician's breakdown risk threshold τ.

Comparing the distributions reveals two differences in effective cost conditions across the

Figure 6: Distribution of the cost ratio τ



Notes: This figure shows the distribution of the threshold $\tau(v_{it})$ for the pre period (blue) and the post period (red). Differences in the distributions for the two periods arise from (1) differences in the estimates of the cost parameters γ_0, γ_1, (2) differences in θ, the scale parameter for idiosyncratic costs ϵ, and (3) differences in the dynamic term $\Delta E \log p(w_{it}, x_{it}; v)$.

periods. First, on average, the effective cost ratio in the post period (0.115) is greater than that in the pre period (0.091). Second, the standard deviation of the effective cost ratio is substantially higher in the post period (0.102) than in the pre period (0.072). This difference is driven by a difference in the estimates of the scale parameter θ for the distribution of idiosyncratic costs: $\theta_{\text{pre}} = 25.4$ and $\theta_{\text{post}} = 17.9$. Since the standard deviation of the logistic distribution is $\frac{\pi}{\sqrt{3}}\frac{1}{\theta}$, this means that, in the post period, there is $\frac{\theta_{\text{pre}} - \theta_{\text{post}}}{\theta_{\text{post}}} \approx 42\%$ more variation in τ conditional on (w, x) than there is in the pre period.

Although these differences in estimated effective costs are dramatic, they are—in context—not altogether surprising. Between the end of the pre period (February 2020) and the beginning of the post period (March 2021), the PFC fleet—like the rest of the world—experienced the multifarious effects of a global pandemic. Among the most salient of these effects for the fleet were changes in the availability of parts. In the context of our model, a part essential to a repair being unavailable would be represented by the effective cost of the repair being extremely high (i.e., getting a very high draw of ϵ).

8 Counterfactual analysis: The value of PredictFix

The previous section showed that when technicians are armed with PredictFix, their behavior reflects a superior ability to predict breakdown risk. In this section, we *quantify* the value of the change in technician behavior induced by PredictFix. Unlike the reduced-form evidence presented in Section 4, these calculations use our cost estimates to control for differences in cost conditions in the pre and post periods.

To quantify the value of PredictFix, we simulate repair and breakdown histories for three scenarios. These scenarios parallel the three ROC curves presented in the previous section. In the first scenario, the technician's predictions of breakdown risk are given by $\boldsymbol{\rho}_{\text{pre}}$ (*without* PredictFix); in the second, the technician's predictions of breakdown risk are given by $\boldsymbol{\rho}_{\text{post}}$ (*with* PredictFix); in the third scenario, again included as a benchmark of *optimal* behavior, the technician's predictions of breakdown risk are given by $\boldsymbol{\pi}$ (the objective breakdown risk). In all three scenarios, we impose the cost conditions of the pre period, as captured by our pre-period estimates of θ and γ.[34] The first scenario, therefore, is a simulated version of the *factual* pre period, while the second scenario is a *counterfactual* where technicians get access to PredictFix while the cost conditions are held constant. Thus, the difference in outcomes between these two scenarios speaks to the value of PredictFix as reflected in its effect on technician behavior.

8.1 Preliminary steps: Estimating the transition process

Evaluation of these counterfactuals requires placing additional structure on the transition process. Recall that the approaches to identification and estimation of the dynamic model presented in Sections 5 and 6 are tailored to one of our setting's main challenges: the high dimensionality of the state variable. Accurately estimating transition processes in such a setting is infeasible without strong functional form assumptions. We avoided making such assumptions by developing an approach to estimation that required only minimal assumptions on and knowledge of the transition process. This minimal approach to the transition process, however, is not sufficient for evaluating dynamic counterfactuals, which requires simulating draws of the state in period $t+1$ given the state and technician's action in period t. Therefore, before proceeding, we must put some additional structure on the state transitions.

To overcome the challenge of estimating the transition process for the high-dimensional state variable (w, x), we instead estimate and simulate transitions in terms of a lower-dimensional object. Recall that the two objects that determine the technician's perceived static payoffs are the cost ratio and the beliefs about breakdown risk. We treat each of these as a component of

[34] We can also do this exercise using post-period cost estimates; doing so produces results that are qualitatively very similar.

a two-dimensional state variable, which we denote (g_{it}, υ_{it}), where

$$g_{it} = g(w_{it}) = \gamma_0 + w_{it}'\gamma_1,$$
$$\upsilon_{it} = \Lambda^{-1}(\rho(x_{it})) = \lambda_0 + x_{it}'\lambda_1,$$

$g_{it} \in \mathbb{R}$ is the average of the cost ratio distribution given w_{it}, and $\upsilon_{it} \in \mathbb{R}$ is a transformed version the perceived breakdown risk, respectively.

We model the transitions of g and υ as independent Markov processes:

Assumption 6 (Transition processes). *g_{it} and υ_{it} evolve as independent Markov processes. For υ_{it}, this means*

$$\upsilon_{it+1} \mid \upsilon_{it} \sim F^{\upsilon}(\cdot\, ;\, \upsilon_{it}, a_{it})$$

Moreover, the transition process for g_{it} is AR(1):

$$g_{it+1} = \beta_0^g + \beta_1^g g_{it} + \eta_{it+1}^g$$

where $\eta_{it+t}^g \sim iid\, N(0, \sigma^g)$.

Operationalizing this assumed transition process for simulating counterfactual scenarios requires estimating the AR(1) parameters $\beta_0^g, \beta_1^g, \sigma^g$ and estimating the conditional distributions $F_{\text{pre}}^{\upsilon}, F_{\text{post}}^{\upsilon}$. To allow for flexibility in the transition, we estimate $F_{\text{pre}}^{\upsilon}, F_{\text{post}}^{\upsilon}$ using Gaussian Mixture Regression, as described by Sung (2004).[35] For details, see Appendix D.6.

In addition to allowing us to simulate drawing states from the transition process, obtaining this estimate of the transition process is also necessary to compute counterfactual ex-ante expected value functions. Recall from Section 5 that the dynamic component of the technician's payoff is given by the difference in ex-ante expected value functions, $EV_1(w,x) - EV_0(w,x)$, which captures the difference in expected future payoffs conditional on doing versus not doing a repair in the current period. Our approach to estimation leveraged the insight that, under Assumption 5, this difference can be written in terms $\Delta Ep(w,x)$, which (expect for the parameter ν) can be estimated directly from technician choice data. While that approach works well in cases where one *observes* technicians' choices, it does not work for *counterfactual* scenarios, where technicians' choices in the scenario of interest are naturally not observed. For our counterfactuals, therefore, we must resort to an older approach: computing the ex-ante expected value functions EV_0 and EV_1 by solving the Bellman equation, a fixed-point problem

[35] A key advantage of this choice is that the conditional distributions of the next period's υ is a Gaussian mixture. This means that expectations over this variable can be evaluated using Guassian quadrature.

in the function space. Then, having computed these functions, we know all the elements of the technician's payoffs and can therefore simulate choices.[36]

The goal of our counterfactual analysis is to quantify the costs—both repair costs and breakdown costs—that result from the technician's choices in each counterfactuals. Computing breakdown costs requires simulating breakdown outcomes. To do this, we make use of the estimates of the logistic regression (11) presented in Table 7, which capture the probability of breakdown conditional on $\upsilon = \Lambda^{-1}\left(\boldsymbol{\rho}_j\right)$ for $j =$ pre, post.

Finally, to make the results of our counterfactuals more interpretable, we want to report costs in dollar terms. However, recall that only the *ratio* of costs τ is identified from the data. To convert this to dollars, we choose a value for the breakdown cost. From conversations with contacts at PFC, we understand that a reasonable value for B would be in the range of \$5,000 - \$10,000.[37] We choose the value $B =$\$5,000 at the low end of that range. If we instead chose a value of $B =$\$10,000 at the high end of the range, this would simply scale all dollar-denominated counterfactual results by a factor of 2.

8.2 Counterfactual results: The value of PredictFix

With our estimates of cost, perceived breakdown risk, objective breakdown risk, transition processes, and ex-ante expected value functions, we can simulate histories of technician decisions for our three scenarios. In all three scenarios, the cost conditions are those of the pre period, but the scenarios differ in technicians' perceived breakdown risk. For the three scenarios, the perceived breakdown risk is $\boldsymbol{\rho}_{\text{pre}}, \boldsymbol{\rho}_{\text{post}}$, and $\boldsymbol{\pi}$, respectively. For each, we simulate 1 million truck-years (52 million truck-weeks). From these simulations, we can compute the frequency of repairs and breakdowns in each scenario. Additionally, since we know the repair costs and breakdown costs, we can calculate the costs incurred under each scenario.

The simulation results for our three scenarios are presented in Figure 7. In this figure, the average annual repair costs (on the x-axis) and the average annual breakdown costs (on the y-axis) are both normalized so that the values are interpreted *relative* to the optimal scenario where technicians know the true risk of breakdown. The blue point represents the costs incurred by a technician without PredictFix (with beliefs $\boldsymbol{\rho}_{\text{pre}}$), and the red point represents the costs incurred by a technician with PredictFix (with beliefs $\boldsymbol{\rho}_{\text{post}}$). Naturally, we think of the technician's objective as minimizing total costs (summing repair costs and breakdown costs);

[36] In keeping with the assumption that the idiosyncratic cost shock is iid logistic, we draw $\epsilon_{it} \sim \text{Logistic}\left(\theta_{\text{pre}}\right)$ for each truck-week.

[37] Recall that these figures include not only the costs of labor, materials, and towing, but also the substantial opportunity costs and risk of damage to relationships with customers and drivers that are incurred with an unanticipated breakdown.

the dashed diagonal lines represent isocost curves, along which the total cost incurred is constant.

The technician with PredictFix resides on a lower isocost curve than the technician without PredictFix; in particular, the former's average annual costs are \$343 less than the latter. We interpret this difference as the *value of PredictFix*, equal to \$343 (or \$686 if the higher breakdown cost were chosen) per truck per year. Comparing the technician with PredictFix to the technician who knows the true risk of breakdown, the latter has costs that are only about \$19 lower per truck per year. Thus, we can say that PredictFix enables technicians to achieve $\frac{343}{343+19} = 94.8\%$ of all cost savings that could be achieved by improving the quality of decision-making. With PredictFix, technicians achieve a quality of decision-making that is very nearly optimal.

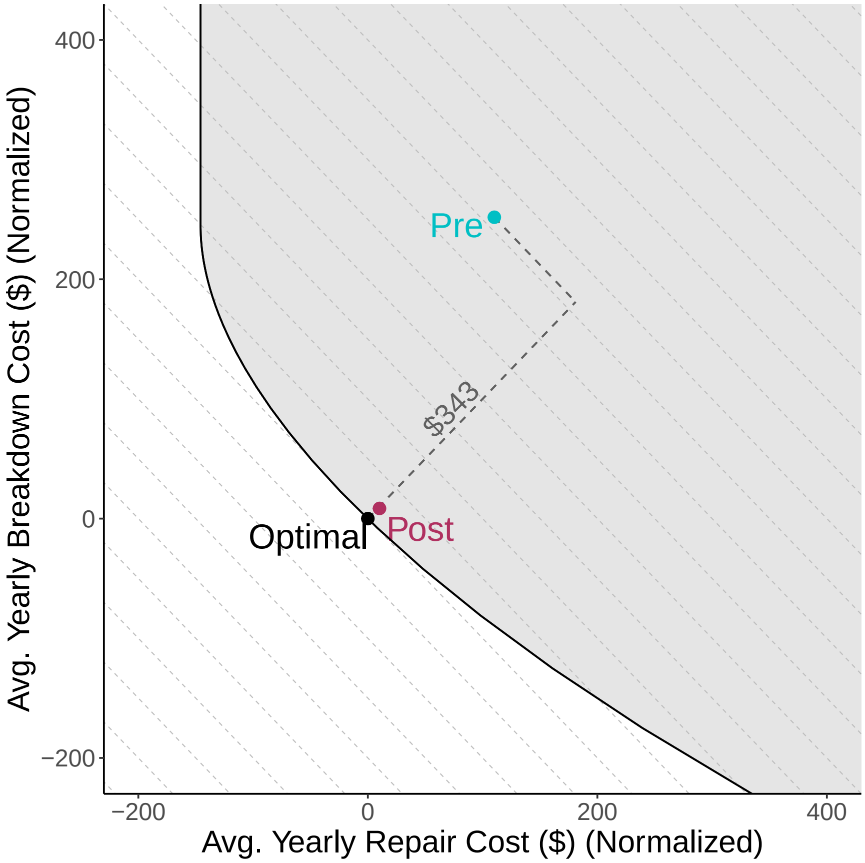
Looking at the relative locations of the Pre and Post points in Figure 7 can help us understand how these cost savings are achieved. Most of these savings—\$243, or 71%—come from a reduction in the frequency of breakdowns, as indicated by the vertical difference between the Pre and Post points. The remaining \$100, or 29%, of the cost savings comes from a reduction in repair costs, as indicated by the horizontal difference between the Pre and Post points. Interestingly, this is not the result of the technician with PredictFix doing fewer repairs—on the contrary, the technician with PredictFix does about 2.5 more repairs per year than the one without. Rather, it is the result of a difference in *which* repairs the technician does. Without PredictFix, the technician is more prone to greatly overestimating the risk of breakdown; she overestimates the risk of breakdown by 5% or more for nearly 1% of all truck-weeks. The technician with PredictFix almost never (for only 0.05% of truck weeks) makes such large prediction errors. This tendency to greatly overestimate the risk of breakdown leads the technician without PredictFix to perform unnecessary repairs when doing so is quite costly.

To dig deeper into these findings, we examine technician mistakes, i.e., decisions made by technicians that deviate from those of the optimal technician who knows the true probability of breakdown. Each panel of Figure 8 plots in gray the distribution of the difference between the expected breakdown cost absent repair and the cost of a repair (including the dynamic effects of the repair):

$$\pi(x)B - \left(\gamma_0 + w'\gamma_1\right) + \delta\left[EV_1^{\pi}(w,x) - EV_0^{\pi}(w,x)\right] + \epsilon$$

It is optimal to do a repair if and only if this difference is greater than zero. Thus, observations that fall in the part of the distribution to the right of zero would, optimally, be repairs; those to the left of zero would, optimally, be non-repairs. To visualize the way in which actual technicians' decisions differ from the optimal decisions, we plot in yellow the distribution of

Figure 7: Repair and breakdown costs with and without PredictFix



Notes: Points indicate the average annual costs incurred from repairs and breakdowns in each decision-making scenario. Repair costs are inclusive of ϵ cost shocks, which account for the negative average repair costs. The colored lines (which have slope -1) are isocost curves. The black curve represents the production possibilities frontier, and the region above the PPF (shaded gray) is the feasible region. To construct this PPF, we simulate the decisions of a technician who knows π and places weight ζ on breakdown costs relative to repair costs for each $\zeta \in (0, \infty)$. Each point on the PPF is parameterized by a different value of ζ. Naturally, lower iso-cost curves are better.

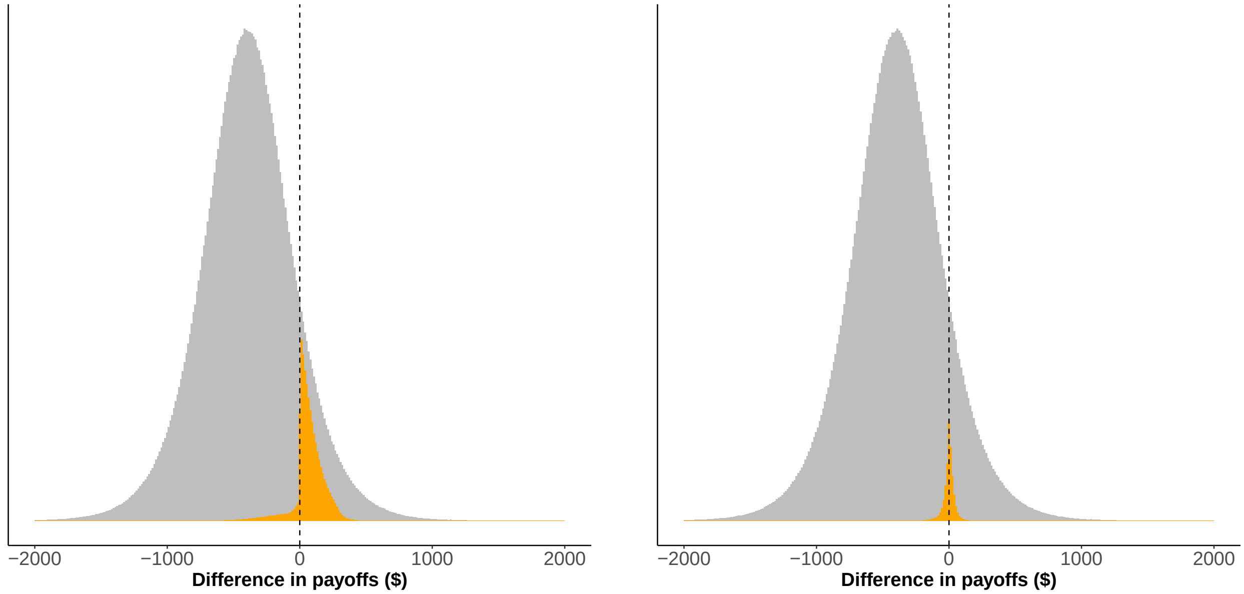
the difference in payoffs for the set of simulated observations where technicians make mistakes, i.e., choose different actions than the optimal technician. In this figure, the ratio of the yellow area to the gray area captures the frequency of such mistakes.

Comparing panels (a) and (b), we can see that the technician without PredictFix makes mistakes more frequently than the technician with PredictFix. The former makes mistakes for 6.5% of all truck weeks, compared with 1.5% for the latter. In addition to greatly reducing the frequency of technician mistakes, PredictFix also dramatically changes the composition of mistakes, i.e., the proportion that are false positives (unnecessary repairs, to the left of zero) versus false negatives (failures to do optimal repairs, to the right of zero). For the technician without

Figure 8: Mistakes and their costliness

(a) Without PredictFix (ρ_{pre})

(b) With PredictFix (ρ_{post})



Notes: Each figure shows the distribution of $\pi(x)B-(\gamma_0+w'\gamma_1)+\delta\left[EV_1^{\pi}(w,x)-EV_0^{\pi}(w,x)\right]+\epsilon$, the difference between the (inclusive) payoff from doing a repair and not doing a repair. Plotted in gray is the distribution of this difference for all observations. Plotted in yellow is the distribution of this difference for the set of observations where the technician makes a "mistake", i.e. makes a choice different from the optimal choice, under each set of beliefs. The frequency of mistakes under each set of beliefs is represented by the ratio of the yellow area to the gray area.

PredictFix, false negatives account for nearly 90% of mistakes. For the technician with PredictFix, on the other hand, false positives and false negatives occur with approximately equal frequency. The fact that PredictFix enables technicians to eliminate so many false negatives explains why about 70% of PredictFix-driven cost savings come from a decrease in breakdown costs.

9 Conclusion

This paper studies how human decision-makers use predictive AI tools. We explore this issue using detailed data on repair decision-making from a large fleet of heavy-duty trucks. In early 2020, the technicians charged with the maintenance of these trucks were given access to PredictFix, a high-quality AI tool designed to predict truck breakdowns. We study how technician decision-making changes with the introduction of this tool. Our analysis draws on rich data that provide insight into every aspect of the technician's decision-problem: we observe the universe of repair decisions made by technicians, the timing of AI-generated alerts, and the minutely detailed, high-frequency data generated by each truck's network of sensors.

In a descriptive comparison of technician decisions and outcomes before versus after PredictFix's introduction, we find that technicians exhibit a statistically significant response to PredictFix alerts; however, we see no resultant change in key fleet outcomes (repair and breakdown frequency), a finding likely affected by pandemic-induced changes in costs concurrent with PredictFix's introduction.

To quantify the effects of PredictFix while flexibly controlling for changes in costs, we develop a dynamic structural model of technician repair decision-making. In bringing this model to the data, we build upon the insights of Hotz and Miller (1993) to develop an approach to estimating dynamic discrete choice models with high-dimensional state variables.

Using this method to estimate our model, we find that technicians' decisions with PredictFix reflect a substantially improved ability to predict breakdowns. Holding cost conditions fixed, the introduction of PredictFix enables technicians to reduce realized costs by $343-$686 per truck per year; equivalently, PredictFix narrows the cost gap between actual decision-making and optimal decision-making by about 95%.

While this paper addresses some key questions about how human decision-makers use AI tools, other important questions might be addressed by future research. First, it would be useful to evaluate heterogeneity in the way human agents use predictive AI, particularly since such heterogeneity has potentially significant implications for the labor market in a world where AI is widespread. Although this question is beyond the scope of the current paper, it is one that we hope to pursue in future. With data on technician characteristics, we could ask how the effects of PredictFix on decision-making differ with, for instance, technicians' tenure or experience. Second, while our analysis acknowledges and allows for incentive and agency problems within the PFC organization, these issues are not the focus of our paper. We do not attempt, for instance, to estimate the effect of PredictFix on incentive misalignment or technician risk aversion. Further exploration of the interaction between AI, decision-making, and principal-agent dynamics would be worthwhile.

While important future questions remain, our paper represents a key step toward understanding the real-world effects of predictive AI on important economic decisions made by humans. In many domains, AI's ability to predict payoff-relevant variables far surpasses human ability. Yet, in most domains, human judgment still plays a significant role. While recent advances in AI have undoubtedly been stunning, there are few settings in which firms have opted for full automation of important economic decisions. As long as humans remain in the decision-making loop, the way they—with all their biases and behavioral quirks—interact with AI systems will continue to determine how this technology shapes economic outcomes.

References

ABALUCK, J., L. AGHA, D. C. CHAN JR, D. SINGER, AND D. ZHU (2020): "Fixing misallocation with guidelines: Awareness vs. adherence," Tech. rep., National Bureau of Economic Research.

AGARWAL, N., A. MOEHRING, P. RAJPURKAR, AND T. SALZ (2023): "Combining human expertise with artificial intelligence: Experimental evidence from radiology," *Working paper*.

ALBRIGHT, A. (2023): "The hidden effects of algorithmic recommendations," *Preprint. Last accessed March*, 28, 2023.

ANGELOVA, V., W. DOBBIE, AND C. YANG (2022): "Algorithmic Recommendations and Human Discretion," *Working Paper*.

ARCIDIACONO, P. AND P. B. ELLICKSON (2011): "Practical methods for estimation of dynamic discrete choice models," *Annu. Rev. Econ.*, 3, 363–394.

ARCIDIACONO, P. AND R. A. MILLER (2011): "Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity," *Econometrica*, 79, 1823–1867.

ARMITAGE, S., F. PINTER, AND R. YANG (2023): "Roadside Infrastructure, Parking, and Electric Trucks," .

BAKER, G. P. AND T. N. HUBBARD (2003): "Make versus buy in trucking: Asset ownership, job design, and information," *American Economic Review*, 93, 551–572.

——— (2004): "Contractibility and asset ownership: On-board computers and governance in US trucking," *The Quarterly Journal of Economics*, 119, 1443–1479.

CHANDRA, A. AND D. O. STAIGER (2007): "Productivity spillovers in health care: evidence from the treatment of heart attacks," *Journal of political Economy*, 115, 103–140.

CHEN, T., T. HE, M. BENESTY, AND V. KHOTILOVICH (2019): "Package 'xgboost'," *R version*, 90, 1–66.

CHEN, T., T. HE, M. BENESTY, V. KHOTILOVICH, Y. TANG, H. CHO, K. CHEN, R. MITCHELL, I. CANO, T. ZHOU, ET AL. (2015): "Xgboost: extreme gradient boosting," *R package version 0.4-2*, 1, 1–4.

CURRIE, J. M. AND W. B. MACLEOD (2020): "Understanding doctor decision making: The case of depression treatment," *Econometrica*, 88, 847–878.

De-Arteaga, M., R. Fogliato, and A. Chouldechova (2020): "A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society: series B (methodological)*, 39, 1–22.

Friedman, J. H. (2001): "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, 1189–1232.

Gort, M. and N. Sung (1999): "Competition and productivity growth: The case of the US telephone industry," *Economic Inquiry*, 37, 678–691.

Harris, A. and T. M. A. Nguyen (2021): "Long-Term Relationships in the US Truckload Freight Industry," .

——— (2022): "Long-term Relationships and the Spot Market: Evidence from US Trucking," .

Hendel, I. and Y. Spiegel (2014): "Small steps for workers, a giant leap for productivity," *American Economic Journal: Applied Economics*, 6, 73–90.

Hotz, V. J. and R. A. Miller (1993): "Conditional choice probabilities and the estimation of dynamic models," *The Review of Economic Studies*, 60, 497–529.

Hubbard, T. N. (2000): "The demand for monitoring technologies: The case of trucking," *The Quarterly Journal of Economics*, 115, 533–560.

——— (2001): "Contractual form and market thickness in trucking," *RAND Journal of Economics*, 369–386.

——— (2003): "Information, decisions, and productivity: On-board computers and capacity utilization in trucking," *American Economic Review*, 93, 1328–1353.

Irvin, J., P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. (2019): "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 590–597.

Jakubik, J., J. Schöffer, V. Hoge, M. Vössing, and N. Kühl (2022): "An Empirical Evaluation of Predicted Outcomes as Explanations in Human-AI Decision-Making," .

JALALIFAR, S. A., H. SOLIMAN, A. SAHGAL, AND A. SADEGHI-NAINI (2022): "A self-attention-guided 3D deep residual network with big transfer to predict local failure in brain metastasis after radiotherapy using multi-channel MRI," *IEEE Journal of Translational Engineering in Health and Medicine*, 11, 13–22.

KIM, H.-E., H. H. KIM, B.-K. HAN, K. H. KIM, K. HAN, H. NAM, E. H. LEE, AND E.-K. KIM (2020): "Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study," *Lancet Digit Health*, 2, e138–e148.

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

NICULESCU-MIZIL, A. AND R. CARUANA (2005): "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd international conference on Machine learning*, 625–632.

REVERBERI, C., T. RIGON, A. SOLARI, C. HASSAN, P. CHERUBINI, GI GENIUS CADx STUDY GROUP, AND A. CHERUBINI (2022): "Experimental evidence of effective human-AI collaboration in medical decision-making," *Sci. Rep.*, 12, 14952.

ROSE, N. L. (1985): "The incidence of regulatory rents in the motor carrier industry," *The RAND Journal of Economics*, 299–318.

——— (1987): "Labor rent sharing and regulation: Evidence from the trucking industry," *Journal of Political Economy*, 95, 1146–1178.

RUST, J. (1987): "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher," *Econometrica: Journal of the Econometric Society*, 999–1033.

STEVENSON, M. T. AND J. L. DOLEAC (2022): "Algorithmic risk assessment in the hands of humans," *Available at SSRN 3489440*.

SUNG, H. G. (2004): *Gaussian mixture regression and classification*, Rice University.

TSCHANDL, P., C. RINNER, Z. APALLA, G. ARGENZIANO, N. CODELLA, A. HALPERN, M. JANDA, A. LALLAS, C. LONGO, J. MALVEHY, J. PAOLI, S. PUIG, C. ROSENDAHL, H. P. SOYER, I. ZALAUDEK, AND H. KITTLER (2020): "Human-computer collaboration for skin cancer recognition," *Nat. Med.*, 26, 1229–1234.

VAN BINSBERGEN, J. H., X. HAN, AND A. LOPEZ-LIRA (2023): "Man versus machine learning: The term structure of earnings expectations and conditional biases," *The Review of Financial Studies*, 36, 2361–2396.

YANG, R. (2022): "(Don't) Take Me Home: Home Preference and the Effect of Self-Driving Trucks on Interstate Trade," .

ZOLAS, N., Z. KROFF, E. BRYNJOLFSSON, K. MCELHERAN, D. N. BEEDE, C. BUFFINGTON, N. GOLDSCHLAG, L. FOSTER, AND E. DINLERSOZ (2021): "Advanced technologies adoption and use by us firms: Evidence from the annual business survey," Tech. rep., National Bureau of Economic Research.

A Setting and data appendix

Technician incentives and agency issues As repair decisions are made by the technicians—agents of PFC—it is natural to ask whether technicians' objectives are aligned with PFC's objective. PFC is aware of, and has taken steps to ameliorate, potential agency issues. Like other private fleets, PFC has historically granted awards to high-performing technicians based on vehicle outcome measures.[38] Yet it is possible that these incentive schemes are not perfectly effective. Residual agency issues may persist. Moreover, it is possible that the severity of these agency issues is altered by the introduction of PredictFix.[39] In addition, it is possible that a misalignment of objectives could arise simply because the process determining the firm's repair and breakdown costs is complicated and not fully understood by technicians. If this is the case, then technicians' behavior might reflect their *perceived* costs, rather than the firm's actual costs.

While questions about principal-agent problems and technicians' objectives are not the focus of this paper, our analysis nevertheless endeavors to address these issues head on. In our discussion of identification in Section 5, we show that our model is identified without making any assumptions about technician preferences or alignment of objectives. This result means that we can recover and analyze our key object of interest, the technician's perceived breakdown risk function $\boldsymbol{\rho}$, regardless of these features of technician behavior. These features do, however, affect the interpretation of our counterfactual analysis, something we discuss further in Section 8.

Unit of analysis: Truck-week Both our reduced-form analysis (Section 4) and our estimation of the structural model (Section 6) require us to discretize time. The two natural period lengths

[38] We do not have access to detailed information about these incentive programs.

[39] For instance, consider a technician without PredictFix does too many repairs because risk aversion leads her to be overly concerned about preventing breakdowns. Once this technician has PredictFix, she might feel that she need not be as risk averse, since if PredictFix does not produce an alert, this gives her "cover" against criticism in the case of a breakdown.

we considered were a day and a week.

We decided to define a period as a week rather than a day primarily because of the low frequency of breakdowns. As shown in Table 2, about 1.5% of truck-weeks have an engine breakdown. If we instead analyzed the data at the daily level, less than 0.25% of truck-days have an engine breakdown. The latter is such a low base rate that it would have presented challenges. In particular, obtaining a high-quality measure of $\pi(x)$—which we estimate by training a GBDT model—would have been infeasible, as most ML classification models (including GBDTs) perform poorly when the data is so extremely imbalanced.

Data cleaning: Avoiding reverse causality Yet aggregating at the weekly level also presents a challenge: concerns about reverse causality. When we aggregate all seven days of a week, x_t reflect information that occurred on seventh day of the week, while $a_t = 1$ might indicate that the technician decided to do a repair on the first day of the week. Suppose that, in the course of doing repairs, technicians always took an action that triggered a particular fault code, which was included in x. This would tend to bias our estimates of beliefs: Since this particular fault code would almost always be accompanied by a repair, the model estimates would indicate that, when technicians saw this this particular fault code, they interpreted it as an indicator of high breakdown risk. In fact, the causality would go in the opposite direction, as it is the technician's action that causes the fault.

One surefire way to avoid this problem is to not include any same-week data in x, i.e., have x_t only reflect fault and sensor data from weeks $t-1, t-2, \ldots$. However, this would introduce other biases: if technicians actually respond to real-time rather than past-week data, this strategy would indicate an erroneously low degree of technician responsiveness to x_t.

We instead choose a less extreme approach: We begin with a vector x_t that includes both same-week and past-week fault and sensor information. We then train—as described in Section D.1.1—a GBDT model, using (w_t, x_t) to predict a_t, whether a repair is done in week t. We then inspect the set of variables with the highest "gain," i.e., those that contribute the most to the GBDT model's predictions. Using institutional knowledge (and a bit of common sense), we identify variables that fall into one of two suspect categories: (i) the variable's gain is disproportional to any genuine information about breakdown risk it could possibly contain, and (ii) the variable seems likely to be affected by actions taken by a technician in the course of a repair.

We identify one set of variables that clearly falls into both categories: fault codes referring to the telematics system or telematics connectivity. Several such fault codes are among the variables with the highest gain; moreover, it seems plausible that certain classes of repairs involve the telematics system being disconnected from its power source.

In addition to these telematics fault codes, we identify several sensors that fall into categories (i) or (ii). It is slightly more difficult to evaluate these sensor variables, as it is not easy to identify what patterns in these variables the GBDT model (which is not easily interpretable) is picking up on. However, several variables with high gain refer to, for instance, the maximum weekly recorded level of a coolant or exhaust fluid. It seems plausible that this is the result of technicians "topping off" these fluids while doing repairs.

Due to these concerns, our model estimation and training of GBDTs excludes from x_t all variables referring to week-t telematics-related faults and all variables referring to week-t sensor measurements. While the latter is perhaps overly restrictive, we feel that it is warranted, as any bias resulting from reverse causality would mean that our estimates $\boldsymbol{\rho}$ would no longer have the desired interpretation.

B Appendix for Section 4

B.1 A primer on ROC curves

For any continuous predictor (in our case, $\hat{\pi}(x)$) of a binary outcome y (in our case, an engine breakdown), the ROC curve gives a visual representation of the predictor's quality by plotting two measures of quality for each of the various *binary* predictors that could be formed using the continuous predictor.

For each threshold $\eta \in [0,1]$, we can form a binary predictor:

$$b_i^\eta = \mathbb{1}\{\hat{\pi}_i > \eta\}$$

where $b_i^\eta = 1$ means that we predict a "positive outcome" ($y = 1$; in our case, a breakdown) for observation i and $b_i^\eta = 0$ means that we predict a negative outcome ($y = 0$; in our case, no breakdown). How well b^η predicts actual outcomes y is captured by two statistics: the True Positive Rate (TPR) and False Positive Rate (FPR). The TPR indicates what proportion of actual positives are (correctly) predicted to be positive, and the FPR indicates the proportion of actual negatives that are (incorrectly) predicted to be positive:

$$\text{TPR}(\eta) = \frac{\text{\# True Positives}}{\text{\# Actual Positives}} = \frac{\sum_i b_i^\eta y_i}{\sum_i y_i}$$

$$\text{FPR}(\eta) = \frac{\text{\# False Positives}}{\text{\# Actual Negatives}} = \frac{\sum_i b_i^\eta (1 - y_i)}{\sum_i (1 - y_i)}$$

In choosing η, one would face a tradeoff between a higher TPR and a lower FPR. As one decreases η (chooses a less conservative threshold for predicting that an observation is “positive”), one may capture more true positives, but also sweep in more false positives. The ROC curve represents a sort of “production possibilities frontier” for binary classifiers by plotting $(\text{FPR}(\eta), \text{TPR}(\eta))$ for all $\eta \in [0,1]$.

As with a standard PPF, a higher curve is better (in this case, better in terms of predictive quality). Therefore, the area under the ROC curve (AUC-ROC or AUC) is a widely used measure of the a predictor’s quality. In addition to this geometric interpretation, the AUC measure also has a useful probability interpretation:

$$\text{AUC} = \Pr\left(\hat{\pi}_{i_1} > \hat{\pi}_{i_0} \mid y_{i_1} = 1, y_{i_0} = 0\right)$$

That is, if one randomly chose a positive observation and a negative observation, the AUC gives the probability that the former has higher predicted $\hat{\pi}$ than the latter.

From both the geometric and probability interpretations, we can see that a perfect (“oracle”) predictor would have AUC $= 1$; on the other hand, a predictor that is no better than random at predicting outcomes would have AUC $= 0.5$ (and would have an ROC curve on the 45-degree line).

B.2 Discussion of PredictFix alert quality

B.3 The role of additional information and technician discretion

Consider a technician who makes optimal repair decisions and has access to PredictFix. Does she always do a repair when a truck has a PredictFix alert? Or does she incorporate additional information contained in the full state of the truck (x) to make her decision?

Additional information contained in x might be valuable to this technician for two reasons: (1) the non-optimality of PredictFix alerts and (2) variation in costs.

(1) Combining x and PredictFix to form an optimal classifier In the analysis above, we showed that PredictFix alerts are very good binary binary classifiers, but still sub-optimal ones. The fact that they lie below—rather than on—the ROC curve (Figure 3) indicates that they do not fully exploit all of the breakdown-relevant information in x. Thus, having access to x itself would allow the optimal technician to form a more informative binary classifier than PredictFix alone.

(2) Variation in cost threshold Setting aside issue (1), let us suppose that PredictFix alerts were an optimal binary classifier, so that they fully exploited the breakdown-relevant information in x. This would mean that, for some threshold π^*,

$$\text{PredictFix}_i \Leftrightarrow \pi(x_i) \geq \pi^*$$

From the model of technician decision-making (Section 5), the optimal technician's decision rule also has a threshold form:

$$a_i = 1 \;\Leftrightarrow\; \pi(x_i) > \frac{\text{Cost of repair}}{\text{Cost of breakdown}} \equiv \tau(v)$$

This implies that whether additional information from x is useful in determining the technician's decision depends on $\tau(v)$. Table 8 summarizes the technician's decisions as a function of PredictFix alerts and the cost threshold. If there is a PredictFix alert and the cost of a repair is relatively low ($\tau(v) \leq \pi^*$), then the technician will do a repair. If there is not a PredictFix alert and the cost of a repair is relatively high ($\tau(v) > \pi^*$), then the technician will not do a repair. In these two cases, the binary PredictFix output contains all of the information the technician would need to decide whether to do a repair. In the other two cases, however, the binary PredictFix output is insufficient. Additional information on breakdown risk contained in x would be useful to decide whether a repair is optimal.

Table 8: Optimal decisions given optimal binary signal

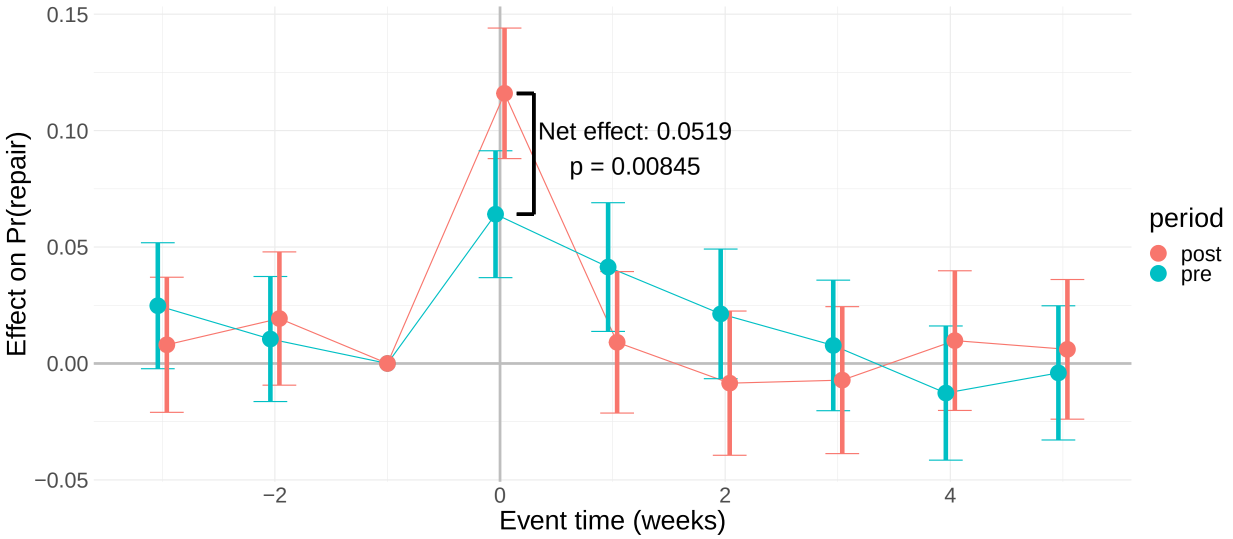
	$\tau(v) \leq \pi^*$	$\tau(v) > \pi^*$
PredictFix alert	Repair	?
No PredictFix alert	?	No repair

This discussion highlights that—even if PredictFix were an optimal binary predictor of breakdowns—it would still be optimal for technicians to exercise *discretion*. Rather than mechanically responding to a PredictFix alert by doing a repair, the technician should combine it with other information on costs and the observable state of the truck to arrive at a repair decision.

B.4 Appendix for Fact 3

B.4.1 Response to medium-severity PredictFix alerts

Figure 9: Event Study: Response of repairs to (predicted) medium-severity PredictFix alerts



Notes: This figure plots the estimated coefficients $\{\beta_\tau^k\}$ from equation (1). β_{-1}^{pre} and β_{-1}^{post} are normalized to zero, though the inclusion of week fixed effects means that differences in the average probability of repair across the pre and post periods are absorbed.

B.4.2 Event study with actual PredictFix alerts

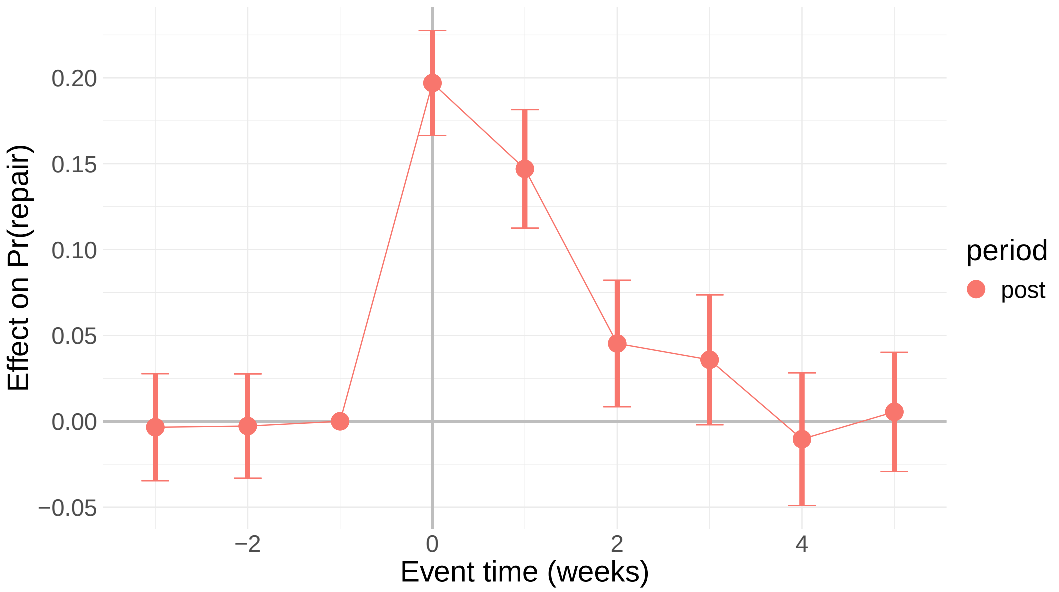
In Section 4.3, we present estimates from an event study estimated on both pre- and post-period data and using predicted PredictFix alerts. An alternative approach would be to estimate an event study using *actual* rather than *predicted* PredictFix alerts and using data only from the post period (the period where actual PredictFix alerts are observed):

$$\text{Repair}_{i,t} = \alpha_0 + \sum_{\tau=-3}^{5} \beta_\tau \text{PredictFix}_{i,t-\tau} + \alpha_i + \gamma_t + \epsilon_{i,t} \tag{12}$$

where $\text{Repair}_{i,t}$ is an indicator for a technician doing an engine repair on truck i in week t and $\text{PredictFix}_{i,t}$ is an indicator for a PredictFix alert for truck i in week t. α_i and γ_t represent truck and week fixed effects, respectively.

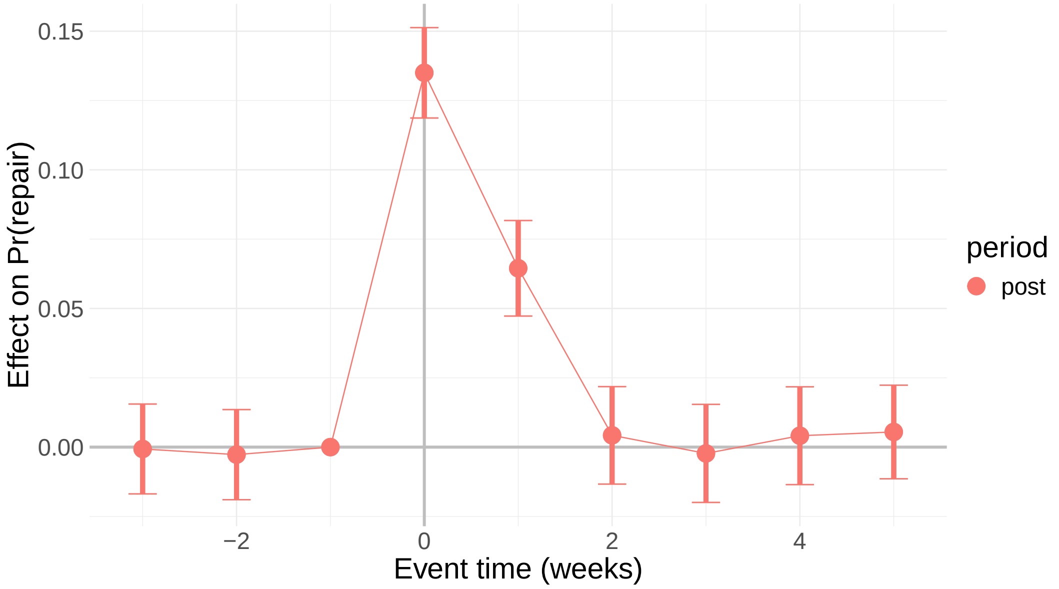
The estimates of coefficients $\{\beta_\tau\}$ are presented in Figures 10 and 11 for high-severity and medium-severity alerts, respectively. In these figures, we see that the probability that a technician does a repair is 19.7pp (13.5pp) higher in the week of a high-severity (medium-severity) PredictFix alert as compared with the week before an alert. However, this is not in itself evidence of technicians responding to PredictFix; it is possible that, instead, this reflects technicians responding to patterns of fault codes and sensor measurements that are correlated with PredictFix alerts. The analysis in the main text and in Appendix B.4.1 uses *predicted*

Figure 10: Event Study: Response of repairs to (actual) high-severity PredictFix alerts



Notes: This figure plots the estimated coefficients $\{\beta_\tau\}$ from equation (12). β_{-1} is normalized to zero. The regression is estimated using only observations from the post period, as there are no PredictFix alerts in the pre period.

Figure 11: Event Study: Response of repairs to (actual) medium-severity PredictFix alerts



Notes: This figure plots the estimated coefficients $\{\beta_\tau\}$ from equation (12). β_{-1} is normalized to zero. The regression is estimated using only observations from the post period, as there are no PredictFix alerts in the pre period.

PredictFix alerts and data from the pre period to tease out responses to alerts versus responses to other signals.

C Model appendix

C.1 Partial identification result

Proposition C.1 (Partial identification of ρ). *Under Assumptions 1-2, ρ is identified up to an affine, increasing transformation. That is, $\tilde{\rho}$ is identified, where $\tilde{\rho}(x) = s(\rho(x))$ for some unknown monotone increasing function $s : [0,1] \to \mathbb{R}$.*

Proof. See Appendix C.2.1. □

This lemma means that, with only Assumptions 1-2, we can recover a function $\tilde{\rho}$ that tells us how technicians *order* states in terms of risk. Since $\tilde{\rho}$ is a monotone increasing transformation of ρ,

$$\tilde{\rho}(x) \geq \tilde{\rho}(x') \iff \rho(x) \geq \rho(x')$$

While we introduce additional assumptions in Section 5, this result implies that a critical part of our analysis—the ROC-based evaluation of how PredictFix affects the technician's ability to predict breakdowns in Section 7.1—is robust to violations of those later assumption.

C.2 Proofs

C.2.1 Proof of Proposition C.1

Proof. First, without loss of generality, we can rewrite

$$g(w) = \gamma_0 + \tilde{g}(w)$$

where $\tilde{g}(w_0) = 0$ for some w_0. Then, $\theta\tilde{g}$ is identified from technicians' responsiveness to cost shifters:

$$\frac{d}{dw} \log \frac{p(x,w)}{1 - p(x,w)} = -\theta \nabla \tilde{g}(w)$$

This gradient, together with the initial condition $\theta\tilde{g}(w_0) = 0$, means that the function $\theta\tilde{g}$ is identified.

Then, since both the log-odds ratio $\log \frac{p(w,x)}{1-p(w,x)}$ and the function $\theta \tilde{g}(w)$ are identified, so is their difference, which we call $\tilde{\boldsymbol{\rho}}(x)$:

$$\tilde{\boldsymbol{\rho}}(x) = -\theta\gamma_0 + \theta \boldsymbol{\rho}(x)$$

Since $\theta > 0$, $\tilde{\boldsymbol{\rho}}$ is an affine, increasing transformation of $\boldsymbol{\rho}$. □

C.2.2 Proof of Proposition 1

Proof. Recall that the proof of Proposition C.1 showed that

$$\tilde{\boldsymbol{\rho}}(x) = -\theta\gamma_0 + \theta \boldsymbol{\rho}(x) \tag{13}$$

is identified. From here, we must prove that, under Assumption 3, we can disentangle this function's constituent parts, separately identifying θ, γ_0, and $\boldsymbol{\rho}$.

Note that the mean and minimum of $\tilde{\boldsymbol{\rho}}(x)$ can be written as

$$\begin{aligned}\mathbb{E}_x \tilde{\boldsymbol{\rho}}(x) &= -\theta\gamma_0 + \theta \mathbb{E}_x \boldsymbol{\rho}(x)\\ \min_x \tilde{\boldsymbol{\rho}}(x) &= -\theta\gamma_0 + \theta \min_x \boldsymbol{\rho}(x)\end{aligned}$$

Writing this in matrix form,

$$\underbrace{\begin{bmatrix} -1 & \mathbb{E}_x \boldsymbol{\rho}(x) \\ -1 & \min_x \boldsymbol{\rho}(x) \end{bmatrix}}_{A} \begin{pmatrix} \theta\gamma_0 \\ \theta \end{pmatrix} = \begin{pmatrix} \mathbb{E}_x \tilde{\boldsymbol{\rho}}(x) \\ \min_x \tilde{\boldsymbol{\rho}}(x) \end{pmatrix} \tag{14}$$

From Assumption 3, $\mathbb{E}_x \boldsymbol{\rho}(x)$ and $\min_x \boldsymbol{\rho}(x)$ are known. Moreover, the matrix A is invertible as long as $\mathbb{E}_x \boldsymbol{\rho}(x) \neq \min_x \boldsymbol{\rho}(x)$, so this system of equations identifies θ and γ_0. It follows that the function $\boldsymbol{\rho}$ is also identified from equation (13). □

C.2.3 Proof of Proposition 2

Technical Condition 1. *There exists some state (w, x) such that*

$$\text{sign}\left(\frac{d^2}{dw^j dx^k} f(w,x)\right) \neq \text{sign}\left(\frac{d}{dw^j} f(w,x) \frac{d}{dx^k} f(w,x)\right)$$

where $f(w_t, x_t) = \mathbb{E}[\log p(w_{t+1}, x_{t+1}) | w_t, x_t, a_t = 0]$, and x^k and w^j are the elements of x and w referred to in Assumption 4, i.e., x^k satisfies Assumption 4(i) and w^j satisfies Assumption 4(ii).

While at first glance, the plausibility of this Technical Condition is perhaps a bit difficult to evaluate, note that this condition need only hold for *one* state $(w, x) \in \mathcal{W} \times \mathcal{X}$. Given that $\mathcal{W} \times \mathcal{X}$ is a very high-dimensional space, it seems very likely that there is some state (w, x) that satisfies the condition.

Proof. **Step 1: Identification of $\mu(\cdot,\cdot,\cdot)$ and ν** Recall from equation (9), that the conditional probability of repair can be written as

$$p(w_t, x_t) = \Lambda\left(-\theta\left[g(w_t) + \rho(x_t)\right] + \delta \Delta E \log p(w_t, x_t)\right)$$

Using Assumption 5 and the facts about the Beta distribution described in Remark **??**, this can be rewritten as

$$p(w_t, x_t) = \Lambda\Big(-\theta\left[g(w_t) + \rho(x_t)\right] + \delta\left[\psi\left(\nu\mu(0, w_t, x_t)\right) - \psi\left(\nu\mu(1, w_t, x_t)\right)\right]\Big) \tag{15}$$

From Remark **??**, we know that

$$\mathbb{E}\mu(a_t, w_t, x_t) = \mathbb{E}\left[p\left(w_{t+1}, x_{t+1}\right) \mid a_t, w_t, x_t\right] = \Pr\left(a_{t+1} = 1 \mid a_t, w_t, x_t\right)$$

The conditional probability on the right-hand side is identified directly from the data.[40] Therefore, the function μ is identified.

Returning to equation (15), let us consider the derivatives of the log-odds ratio $\log \frac{p(w_t, x_t)}{1 - p(w_t, x_t)}$. Under the assumptions of the model, for the jth element of x and the kth element of w_t, the cross-partial comes only from the dynamic terms:

$$\begin{aligned}
\frac{d^2}{dw_t^j dx_t^k} \log \frac{p(w_t, x_t)}{1 - p(w_t, x_t)} &= \delta \frac{d^2}{dw_t^j dx_t^k} \Delta E \log p(w_t, x_t) \\
&= \delta \nu \frac{d}{dw^j}\left[\frac{d}{dx^k}\mu(0, w, x)\psi^{(1)}\left(\nu\mu(0, w, x)\right) - \frac{d}{dx^k}\mu(1, w, x)\psi^{(1)}\left(\nu\mu(1, w, x)\right)\right] \\
&= \delta \nu \Big[\mu^0_{w^j x^k}\psi^{(1)}\left(\nu\mu^0\right) + \nu\mu^0_{w^j}\mu^0_{x^k}\psi^{(2)}\left(\nu\mu^0\right) \\
&\qquad - \mu^1_{w^j x^k}\psi^{(1)}\left(\nu\mu^1\right) - \nu\mu^1_{w^j}\mu^1_{x^k}\psi^{(2)}\left(\nu\mu^1\right)\Big]
\end{aligned}$$

where μ^0_z and μ^1_z denote the derivatives of $\mu(0, w, x)$ and $\mu(1, w, x)$ with respect to some variable z. Letting x^k be the element of x that satisfies Assumption 1(i), i.e., that x^k "resets" after a repair, implies that $\mu^1_{x^k} = 0$ for all (w, x). This means that the last two terms are zero, so the

[40] This is true because, under Logistic distribution assumption, ϵ has full support. So for any state (w_t, x_t), we observe both repairs and non-repairs, and therefore observe choices made in the following periods.

cross-partial becomes

$$\frac{d^2}{dw^j dx^k} \log \frac{p(w,x)}{1-p(w,x)} = \delta \nu \left[\mu^0_{w^j x^k} \psi^{(1)}\left(\nu \mu^0\right) + \nu \mu^0_{w^j} \mu^0_{x^k} \psi^{(2)}\left(\nu \mu^0\right) \right] \tag{16}$$

Everything in this equation except ν is known. We now prove that there is a unique ν that satisfies this equation by proving that the expression on the right-hand side is strictly monotone in ν.

Remark 1 (Properties of the polygamma functions)**.** The so-called "polygamma" functions $\psi, \psi^{(1)}, \psi^{(2)}, \ldots$ have the following properties:

- If k is odd, then $\psi^{(k)}$ is strictly decreasing on $(0, \infty)$.
- If k is even, then $\psi^{(k)}$ is strictly increasing on $(0, \infty)$.

A consequence of these properties is that, since $\mu^0, \mu^1 > 0$, $\nu \psi^{(1)}\left(\nu \mu^0\right)$ is strictly decreasing on $(0, \infty)$ and $\nu^2 \psi^{(2)}\left(\nu \mu^0\right)$ is strictly increasing on $(0, \infty)$.

Now we use make use of the technical condition imposed by Assumption 1(iii): there exists some (w, x) such that

$$\operatorname{sign}\left(\mu^0_{w^j x^k}\right) \neq \operatorname{sign}\left(\mu^0_{w^j} \mu^0_{x^k}\right)$$

If $\mu^0_{w^j x^k} < 0 < \mu^0_{w^j} \mu^0_{x^k}$, then the right-hand side of (16) is *strictly increasing* in ν. If $\mu^0_{w^j x^k} > 0 > \mu^0_{w^j} \mu^0_{x^k}$, then the right-hand side of (16) is *strictly decreasing* in ν. In either case, the right-hand side of (16) is *strictly monotone* in ν. This means that there is a unique value of $\nu \in (0, \infty)$ that satisfies this equation, so this equation identifies ν.

Step 2: Identification of ρ, θ, and γ_0

Now that $\delta \Delta E p(w, x)$ is known, subtracting it from the log odds ratio leaves us with a residual that is a function of x:

$$\begin{aligned} k(w,x) &\equiv \log \frac{p(w,x)}{1-p(w,x)} - \delta \Delta E \log p(w,x) \\ &= \theta \left[-g(w) + \rho(x) \right] \end{aligned}$$

The rest of the proof follows exactly the proof of identification in the static model (see C.2.2 above). □

D Empirical appendix

D.1 Estimation appendix

D.1.1 w and x: Defining and selecting variables

We begin by defining the objects in the data that correspond to the variables w (cost shifters) and x (state of the truck). Next, to make it feasible to take our model to the data, we must first reduce the dimensionality of x, an extremely high-dimensional vector describing the state of the truck. To accomplish this, we use a machine learning-based variable selection procedure.

Table 9: Variable definitions

Model	Data
w: Cost shifters	Capacity variables: - # open work orders at facility - # trucks assigned to facility Other variables: - Facility indicators - Month indicators $\hat{c}_{\text{ML}}$: GBDT-based predictor of tangible repair cost
x: Breakdown predictors	Sensor and fault variables, maintenance history: See Table 1 in Section 3.2.
	Predicted PredictFix variables (from GBDTs): - High priority - Medium priority

Notes: This table describes the full set of potential state variables used in estimation. First, the table presents the set of w variables are potential shifters of repair costs. Second, the table expands the set of variables x relevant to predicting breakdown risk outlined in Table 1. Relative to Table 1, this expanded set includes the two predicted PredictFix variables generated as the output of trained GBDT models (see Section 4.3).

Additional x variables As described in Section 3.2, x includes a wide array of variables derived from truck-generated data (i.e., sensor measurements and fault codes) and the truck's maintenance history. As we bring the model to the data, we add one additional set of variables relating to PredictFix alerts.

Rather than including indicators for actual PredictFix alerts in x, we instead include the *predicted* PredictFix measures described in Section 4.3; in particular, we let x include the two continuous outputs of the GBDTs trained to predict (a) high-priority PredictFix alerts and (b) medium-priority PredictFix alerts. Because actual PredictFix alerts occur only in the post period, this approach allows us to maintain symmetry in estimation between the pre and post periods. Note that, since these predictors are functions of the fault, sensor, and maintenance history variables—which are observable to the technician in both periods—including these predictors in x is both appropriate and valuable in trying to understand how PredictFix changes behavior.

Defining w variables Recall that w represents the set of variables that might affect τ, the ratio of the cost of doing a repair to the cost of a breakdown. As we bring the model to the data, we allow w to include variables related to facility capacity—facility's current number of open work orders and total number of trucks—as well as facility and month indicators. The month indicators are likely to be particularly important given the likely pandemic-induced changes in patterns of demand and availability of parts that occurred between the pre and post periods.

In addition to these cost shifters, we also include in w a term that we call $\hat{c}_{\text{ML}}$, which makes use of the internal accounting costs in PFC's maintenance records set to control for the effect of fault codes and sensor measurements on the *tangible* component of costs. This term is the output of a GBDT regression model trained to predict the tangible cost of a repair based on the truck's fault codes and sensor measurements. The reason we use these *predicted* tangible costs rather than the costs observed in the data are two-fold: First, we observe costs in the data only when a repair actually occurs; in estimating the model, however, we need to know what these costs *would have been* if a repair had occurred. Second, the costs observed in the data are *realized* repair costs, which may differ from the technician's *expected* repair costs at the time she makes the repair decision. The predicted $\hat{c}_{\text{ML}}$ corresponds to the expectation of costs that a technician might form based on observed truck-generated data.

Choosing to include $\hat{c}_{\text{ML}}$ in w means that we will estimate a coefficient (an element of γ) corresponding to this object. The fact that we allow this to be a free parameter implies that an increase of \$1 in $\hat{c}_{\text{ML}}$ may not correspond to a \$1 increase in repair costs. The reasons for this are twofold: First, the costs that we observe in the maintenance records are *internal accounting costs*. From conversations with members of the PFC fleet management team, we understand that these numbers are not necessarily representative of monetary costs. For instance, the maintenance records compute a work order's labor costs using an hourly wage. This wage, however, does not correspond to the actual marginal cost of labor, as (1) technicians are salaried, not hourly, and (2) the wage is adjusted to include various elements of the facility's overhead. Second, perhaps in part due to this difference between actual costs and accounting costs, our ability to predict these costs is relatively poor: the out-of-sample pseudo-R^2 is 0.251.

D.1.2 Variable selection

The number of parameters of our model naturally increases with the dimension of x. Given our sample sizes (4,406 truck-weeks in the pre period and 16,925 truck-weeks in the post period), it is not feasible to estimate the model using the full set of more than 2,000 x variables described above. Therefore, we begin with a variable-selection procedure that limits x to be 20-dimensional.

To help identify the x variables most important in determining repair decisions, we train a GBDT model to predict repairs using the full set of x and w variables as predictors. We then select the twenty x variables to include in x_{model} by taking those with the highest "gain," i.e., those that contribute the most to the GBDT model's prediction. These are the x variables we will use in the estimation of our model. This GBDT model, whose predicted repair probabilities we call $\hat{p}_{\text{ML}}$, alsos serve as a nonparametric benchmark against which we compare the goodness-of-fit of our parametric model (see Appendix D.4).

D.1.3 Estimation of μ, the conditional expectation of next period's repair probability

As described in Section 5.2, one element of the dynamic component of the technician's payoff is captured by the term $\Delta E \log p(w, x)$. This term can be written as a function of two primitives, a constant ν and a function $\mu : 0, 1 \times \mathcal{W} \times \mathcal{X} \to \mathbb{R}^{+}$. From Remark **??**(i), the latter can expressed as:

$$\begin{aligned} \mu(a_t, w_t, x_t) &= \mathbb{E}\left[p_{t+1}\left(w_{t+1}, x_{t+1}\right) \mid a_t, w_t, x_t\right] \\ &= \Pr\left(a_{t+1} \mid a_t, w_t, x_t\right) \end{aligned}$$

Constructing the dynamic term, therefore, requires having an estimate of the expected probability of a repair next week, given this week's choice (a_t) and state (w_t, x_t).

To accomplish this, we once again cross-validate, train, and calibrate a GBDT model, this time to predict a_{t+1} using a_t, w_t, and x_t as predictors.[41] The output of this process is a flexible function that can be interpreted as the desired conditional probability.

D.2 Machine learning

In the course of our analysis, we make use of machine learning models—in particular, gradient-boosted decision trees (GBDTs)—several times to learn the relationships between high-dimensional sets of predictors and binary outcome variables. First, in Section 4.1, we

[41] In recognition of the fact that p may differ between the pre and post periods, we also include a post-period indicator as a predictor in training this GBDT.

describe training a GBDT model to predict breakdowns based on states of the truck x. This trained model serves as the basis for $\hat{\pi}(x)$, our objective measure of breakdown risk, something that we use throughout our reduced-form and structural analysis. Second, in Section 4.3, we describe training a pair of GBDT models, which use the observable state of the truck x to predict PredictFix alerts. We train one model to predict high-priority alerts and another to predict medium-priority alerts. We use binary classifiers formed from these trained GBDTs' output to estimate the event study regression (1) that captures technicians' responsiveness to these predicted alerts in both the pre and post periods. In addition, the continuous PredictFix predictors from these trained GBDTs are included as x-variables in the estimation of the structural model. Finally, we train a GBDT model to predict whether a technician does a repair based on the observable state of the truck x and cost shifters w. The out-of-sample goodness of fit of this GBDT model serves as a benchmark against which we evaluate the fit of our parametric model. (See Figure 12.) This subsection describes the details of how all of these GBDT models are trained.

Implementation To estimate our GBDTs, we use XGBoost, an open-source framework for training regularized GBDTs.(Chen et al., 2015) We make use of this framework using the R package `xgboost`.(Chen et al., 2019)

When predicting binary outcomes (breakdowns, PredictFix alerts, repairs), we train each GBDT to maximize the log-likelihood objective function. When predicting a continuous outcome (realized tangible repair costs), we train the GBDT regression model to minimize the sum of squared errors.

Sample splitting, cross-validation, and hyperparameters When training a flexible machine learning model, a key concern is overfitting. It is easy to train a model that simply "memorizes" the data it is trained on, but is very poor in its ability to generalize to examples not included in the training data. For this reason, we implement several forms of regularization and select the appropriate degree of regularization using *cross-validation*.

At the outset, we randomly split our sample into two groups: 5/7 of the data become the *training set*, while the remaining 2/7 is the *test set*. The test set is not used in any way when training the model; it is simply reserved for evaluating the fit of the model after all training is complete.

The training set is further split into five random sub-samples, which are used for cross-validation. Each of the five is held out in turn, with the fifth is used as a sort of temporary test set. Then, analyzing the out-of-sample fit for each, we average over the five samples and take this as a measure of the generalizability of the fitted model. We repeat this cross-validation procedure many times for a large set of hyperparameters. This procedure allows us to determine the appropriate degree of regularization that resolves the bias-variance tradeoff in a way that results in the best out-of-sample fit.

The set of hyperparameters we consider is as follows:

$$(\eta, \gamma, \text{max depth}) \in \mathcal{H} \times \Gamma \times \mathcal{M}$$

where $H = \{\exp(-10), \exp(-9.5), \exp(-9), \ldots, \exp(-2.5), \exp(-2)\}$, $\Gamma = \{0, 5, 10\}$, and $\mathcal{M} = \{2, 4, 8\}$. η is the learning rate, γ is the minimum loss reduction threshold for further partitioning on a node of the tree, and max depth indicates the maximum allowed depth for any tree. More conservative (i.e., more regularized) models result from lower η values, higher γ values, and smaller max depth values.

Recalibration Despite the fact that we train our GBDTs to maximize the log-likelihood, it is not necessarily appropriate to interpret the GBDT outputs as probabilities. This is because, despite our careful cross-validation, the way we resolve the bias-variance tradeoff is likely imperfect, possibly resulting in some degree of either overfitting or underfitting. Either of these will lead to miscalibration in the model output. This is a well-known problem with certain classes of machine learning models, including GBDTs.(Niculescu-Mizil and Caruana, 2005) The two most common solutions to this problem—isotonic regression and Platt scaling—involve applying a monotone transformation to the model output. We opt for isotonic regression, using the training set to fit a monotone, nonparametric regression of actual outcomes (e.g., breakdowns) on the GBDT output. When we then apply this estimated monotone function to GBDT outputs, the result is a well-calibrated fitted probability.

D.3 Constrained maximum likelihood estimation

We solve the constrained maximum likelihood problem (equation 10) using an the Augmented Lagrangian approach. Letting $\beta = (\theta, \nu, \gamma, \lambda)$ denote the set of parameters to be estimated, this approach makes use of the following augmented Lagrangian function

$$\mathcal{L}(\beta, \lambda, \kappa) = f(\beta) + \lambda' g(\beta) + \frac{\kappa}{2} g(\beta)' g(\beta) \tag{17}$$

where λ is a vector of Lagrange multipliers, κ is a hyperparameter, and $g(\beta)$ represents a vector of the constraint functions:

$$g(\beta)=\begin{bmatrix}\mathbb{E}_x\boldsymbol{\rho}(x)-\mathbb{E}_x\boldsymbol{\pi}(x)\\ \min_x^*\boldsymbol{\rho}(x)-\min_x^*\boldsymbol{\pi}(x)\end{bmatrix} \tag{18}$$

where $\min_x^* f(x)$ denotes the mean of the bottom 0.025 percent of values of $f(x)$ in the distribution of x. This choice reflects the fact that the gradient of this $\min^*$ function is, in practice, better behaved than that of the true minimum function.

We then iterate over two steps:

Algorithm 1: Augmented Lagrangian approach

```
Choose initial parameters $\beta_0$;
while Not converged do
    Primal update: $\beta \leftarrow \arg\min_\beta \mathcal{L}(\beta,\lambda,\kappa)$;
    Dual update: $\lambda \leftarrow \lambda + \kappa g(\beta)$;
end
```

Note that, for any λ,κ, the function $\mathcal{L}(\beta,\lambda,\kappa)$ is nonconcave in β. (This is true event if $\lambda=\vec{0}$ and $\kappa=0$ because our choice probability function nests the non-linear function $\boldsymbol{\rho}$ in another non-linear function Λ.) This means that the primal update step requires solving a non-convex optimization problem. To make it more likely that we converge to the global, rather than a local, optimum in the primal step, we use a momentum-based method with momentum parameter 0.9. For each iteration within the primal update step, the step direction update is that of gradient descent, whose computation is made possible by the fact that the gradient of (17) can be computed analytically.

We experiment with a wide range of values for the hyperparameter κ: $\mathcal{K}=\{2^{3k}$ for $k=1,2,\ldots,12\}$. Ultimately, we select the smallest value in $\mathcal{K}$ that yields an optimum with $|\mathbb{E}_x\boldsymbol{\rho}(x)-\mathbb{E}_x\boldsymbol{\pi}(x)|<10^{-5}$ and $\left|\min_x^*\boldsymbol{\rho}(x)-\min_x^*\boldsymbol{\pi}(x)\right|<10^{-5}$. For the pre period, this is $\kappa=2^{18}$; for the post period, it is $\kappa=2^{24}$.

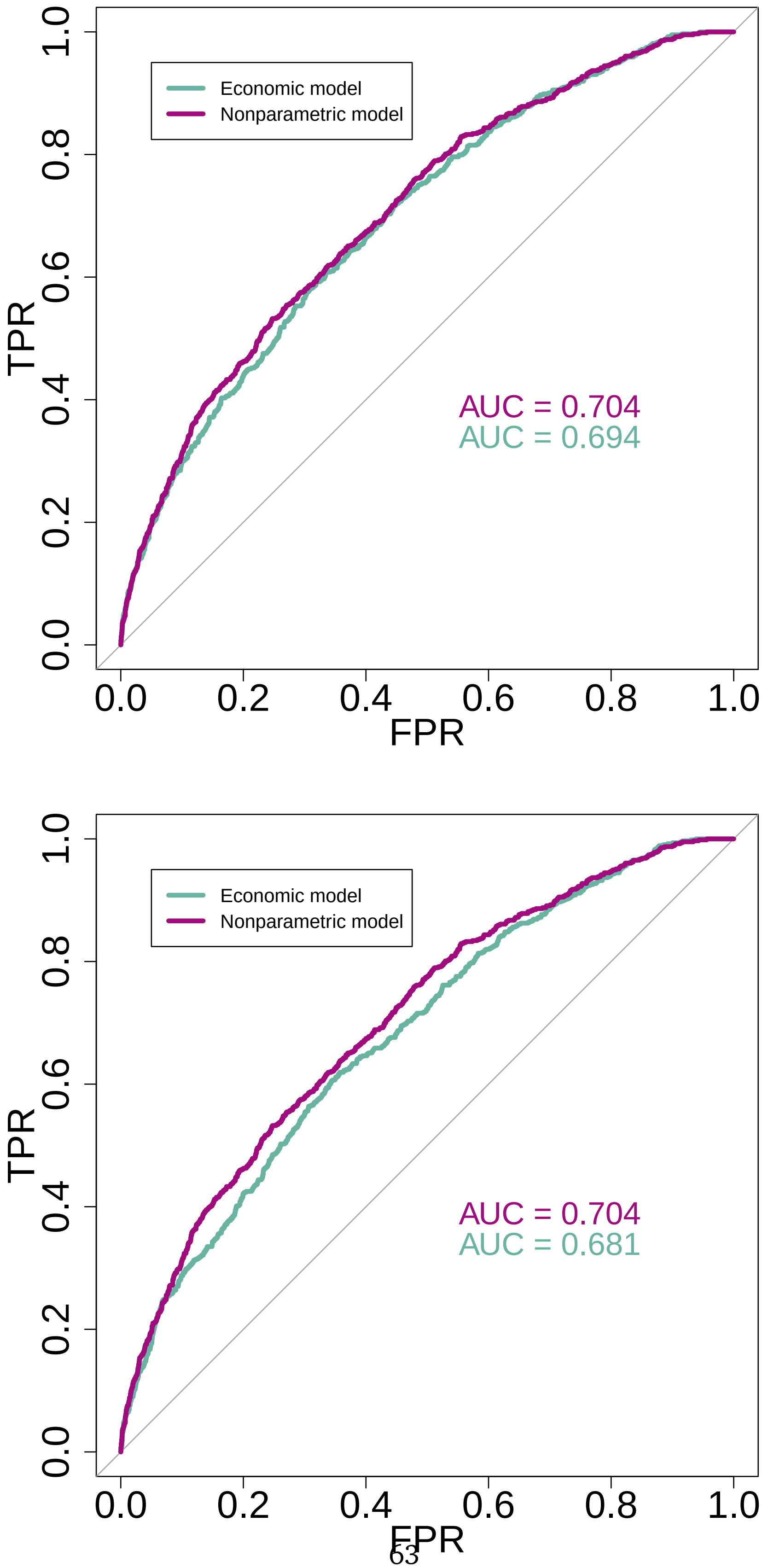
Finally, to speed up estimation, we start with an initial parameter vector β_0 drawn from a favorable region of the part of the parameter space. To find such a region, we use Algorithm 2, which uses logistic regression to estimate a simplified version of our model. The logistic regression in step 5 of the algorithm estimates a model that differs from ours in two respects: (i) it does not have the constraints on beliefs from Assumption 3, and (ii) it has *linear* beliefs, $\alpha_{x0}+x'\alpha_{x1}$. Steps 7-8 adjust the resulting coefficients to account for these differences. This procedure yields a parameter vector β_0 that, in practice, seems to be in the neighborhood of the global optimum β^*. Adding noise to β_0 and using a multi-start procedure helps us overcome the fact that our objective function is non-convex.

```
Algorithm 2: $\beta$ initialization
for $\nu \in H$ do
    Run a logistic regression of repair choices $a_{it}$ on $(w_{it}, x_{it}, \Delta E \log p(w_{it}, x_{it}; \nu)$.;
end
$\nu_0 \leftarrow \arg\min_{\nu \in H} |\text{coefficient on } \Delta E \log p(w_{it}, x_{it}; \nu) - \delta|$ ;
Run a logistic regression of repair choices $a_{it}$ on $(w_{it}, x_{it}, \Delta E \log p(w_{it}, x_{it}; \nu)$. Call the
  resulting vector of coefficients $\hat{\alpha}$.;
$\gamma_1 \leftarrow \hat{\alpha}_w$ (the components of $\hat{\alpha}$ corresponding to $w$);
Adjust $\hat{\alpha}_x$ (the coefficients on $x$) and the constant so that the constraints are
  approximately satisfied by the *linear* beliefs: $\hat{\alpha}_{x0} + x'\hat{\alpha}_x$.;
Convert these linear belief parameters to logistic belief parameters by running a linear
  regression: of $\Lambda^{-1}\left(\hat{\alpha}_{x0} + x'_{it}\hat{\alpha}_x\right)$ on $x_{it}$. The resulting coefficients are our initial guess
  for $\lambda$.;
```

Even with this initialization expedient, estimation is still computationally expensive: Using five-fold parallelization of gradient computation, estimation takes about 20 hours for each value of κ, for each period (pre, post), and for each initial β_0.

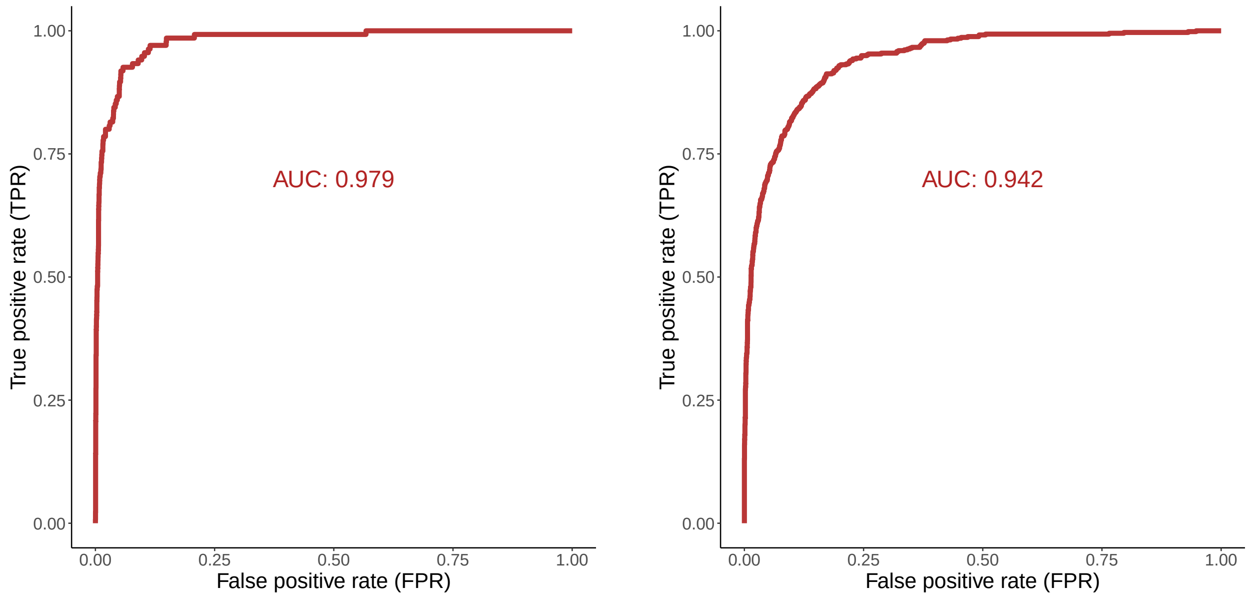
D.4 Goodness of Fit

Figure 12: Model Fit: ROC curves for our model and a nonparametric model



Notes: The green curve is an ROC curve that illustrates the quality of the estimated model's predicted repair probability $p(w, x)$ as a predictor of actual (out-of-sample) repairs. For comparison, the ROC curve for a

Figure 13: ROC curves for ML predictors of PredictFix alerts



(a) High severity alerts

(b) Medium severity alerts

Notes: This figure shows out-of-sample ROC curves illustrating the quality of our GBDT predictors $\widehat{\text{PredictFix}}^{\text{High}}$ and $\widehat{\text{PredictFix}}^{\text{Medium}}$ as predictors as acutal High- and Medium-severity alerts, respecitvely.

D.5 Results appendix

D.5.1 ROC curves and Blackwell informativeness

Proposition D.1. *Suppose that information structures A and B give rise to classifiers with ROC curves characterized by* $\text{TPR}_A(\text{FPR})$ *and* $\text{TPR}_B(\text{FPR})$*. If the ROC curve for A is (weakly) above the ROC curve for B, then information structure A is (weakly) more Blackwell informative than information structure B.*

Proof. Suppose the cost ratio is τ. (For the purposes of this proof, we can, without loss of generality, normalize the cost of a breakdown to 1.) Then, given information structure $i \in \{A, B\}$, the agent's best feasible expected utility for a given FPR is

$$\begin{aligned}\mathbb{E}U_i(\text{FPR}) &= -(1 - \text{TPR}_i(\text{FPR})) - \tau\text{FPR}\\ &= -1 + \text{TPR}_i(\text{FPR}) - \tau\text{FPR}\end{aligned}$$

Let $\text{FPR}_i^* \equiv \text{argmax}_{\text{FPR}} \mathbb{E}U_i(\text{FPR})$ and $\mathbb{E}U_i^* = \mathbb{E}U_i(\text{FPR}_i^*)$. Then,

$$\begin{aligned}
\mathbb{E}U_B^* &= -1 + \text{TPR}_B(\text{FPR}_B^*) - \tau \text{FPR}_B^* \\
&\leq -1 + \text{TPR}_A(\text{FPR}_B^*) - \tau \text{FPR}_B^* \\
&\leq -1 + \text{TPR}_A(\text{FPR}_A^*) - \tau \text{FPR}_A^* \\
&= \mathbb{E}U_A^*
\end{aligned}$$

where the first inequality comes from the fact that A's ROC curve being strictly above B's ROC curve means that, for any FPR, $\text{TPR}_A(\text{FPR}) > \text{TPR}_B(\text{FPR})$. The second inequality comes from the definition of FPR_A^*: it maximizes $\mathbb{E}U_i(\text{FPR})$, so $-1+\text{TPR}_A(\text{FPR}_A^*)-\tau\text{FPR}_A^* \geq -1+\text{TPR}_A(\text{FPR}_B^*)-\tau\text{FPR}_B^*$.

This shows that, for any cost ratio τ, information structure A gives (weakly) higher expected utility. This inequality is strict if the ROC curve for A is strictly above the ROC curve for B at FPR_B^*. Information structure A is therefore (weakly) more Blackwell informative. □

D.5.2 Further evaluation of the effect of PredictFix on prediction quality

While ROC curves are one tool for comparing the quality of predictors, they do not paint a complete picture; in particular, ROC curves only assess predictors' ability to *order* observations by riskiness. The *calibration* of a predictor's values are also important. With this in mind, we additionally compute a goodness-of-fit statistic more widely used in economics which captures both of these dimensions: the likelihood.

For each predictor $\boldsymbol{\rho}_{\text{pre}}, \boldsymbol{\rho}_{\text{post}}$, and $\boldsymbol{\pi}$, we compute the average log-likelihood as follows:

$$\frac{1}{N}\sum_{i=1}^{N}[\text{breakdown}_i \log \boldsymbol{\rho}(x_i) + (1-\text{breakdown}_i)\log(1-\boldsymbol{\rho}(x_i))]$$

and similarly for $\boldsymbol{\pi}$. The results—computed using the same sample restrictions used in the construction of the ROC curves—are presented in Table 10.

These results again show a substantial improvement: the quality of predictions by the technician with PredictFix is closer to the objective benchmark $\boldsymbol{\pi}$ than it is to the the predictions of the technician without PredictFix.

Table 10: Average log-likelihood of breakdown outcomes as predicted by $\boldsymbol{\rho}_{\text{pre}}$, $\boldsymbol{\rho}_{\text{post}}$, and $\boldsymbol{\pi}$

Predictor	Log-likelihood
$\boldsymbol{\rho}_{\text{pre}}$	-0.0751
$\boldsymbol{\rho}_{\text{post}}$	-0.0698
$\boldsymbol{\pi}$	-0.0671

Notes: This table shows the (mean) log-likelihood of observed breakdown outcomes using three different sets of probabilities (i.e., predictors). The first two, $\boldsymbol{\rho}_{\text{pre}}$ and $\boldsymbol{\rho}_{\text{post}}$, are the estimates of technicians' perceived breakdown risk from the model, estimated using pre-period and post-period data, respectively. The third is $\boldsymbol{\pi}$, the objective breakdown probability measure generated as the output of the GBDT model trained to predict breakdowns. For the sample restrictions used in the calculation of each, see Table 6.

D.6 Counterfactuals appendix

Gaussian mixture regression We use the Gaussian mixture regression method of Sung (2004) to estimate four different conditional distributions: $F^{v}_{\text{pre}}|a_t = 0$, $F^{v}_{\text{pre}}|a_t = 1$, $F^{v}_{\text{post}}|a_t = 0$, and $F^{v}_{\text{post}}|a_t = 1$. The estimation procedure for a Gaussian mixture regression of y on x is summarized below:

First, we suppose that the vector (x, y) is distributed as a mixture of M Gaussians, i.e., its pdf is

$$f_{X,Y}(x,y) = \sum_{m=1}^{M} \kappa_m \phi\left(x, y; \mu_m, \Sigma_m\right)$$

where $\{\kappa_m\}$ are weights with $\sum_{m=1}^{M} \kappa_m = 1$. We estimate $\{\kappa_m, \mu_m, \Sigma_m\}$ using the EM algorithm (Dempster et al., 1977).[42]

Then, the distribution of y conditional on an observed value x is

$$f_{Y|X}(y|x) = \sum_{m=1}^{M} \omega_m(x) \phi\left(x; \mu_{mX}, \Sigma_{mX}\right)$$

where the mixing weights $\{\omega_m\}$ are derived using Bayes' Rule:

$$\omega_m(x) = \frac{\kappa_m \phi\left(x; \mu_{mX}, \Sigma_{mX}\right)}{\sum_{m'=1}^{M} \kappa_{m'} \phi\left(x; \mu_{m'X}, \Sigma_{m'X}\right)}$$

[42] To carry out the EM algorithm, we use the Julia package `GaussianMixtures`.