informs
http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 65, No. 3, March 2019, pp. 1136–1150
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Contracting in Medical Equipment Maintenance Services: An Empirical Investigation

Tian Heong Chan,[a] Francis de Véricourt,[b] Omar Besbes[c]

[a] Goizueta Business School, Emory University, Atlanta, Georgia 30322; [b] ESMT European School of Management and Technology, 10178 Berlin, Germany; [c] Graduate School of Business, Columbia University, New York, New York 10027

Contact: tian.chan@emory.edu, http://orcid.org/0000-0002-0286-1854 (THC); francis.devericourt@esmt.org, http://orcid.org/0000-0003-2832-0738 (FdV); ob2105@columbia.edu, http://orcid.org/0000-0002-2982-3794 (OB)

Received: September 4, 2014
Revised: July 9, 2015; August 26, 2016; June 1, 2017
Accepted: August 22, 2017
Published Online in Articles in Advance: June 7, 2018

https://doi.org/10.1287/mnsc.2017.2993

Abstract. Maintenance service plans (MSPs) are contracts for the provision of maintenance by a service provider to an equipment operator. These plans can have different payment structures and risk allocations, which induce various types of incentives for agents in the service chain. How do such structures affect service performance and service chain value? We provide an empirical answer to this question by using unique panel data covering the sales and service records of more than 700 diagnostic body scanners. We exploit the presence of a standard warranty period and employ a matching approach to isolate the incentive effects of MSPs from the confounding effects of endogenous contract selection. We find that moving the equipment operator from a basic, pay-per-service plan to a fixed-fee, full-protection plan not only reduces reliability but also increases equipment service costs. Furthermore, that increase is driven by both the operator and the service provider. Our results indicate that incentive effects arising from MSPs leads to losses in service chain value, and we provide the first evidence that a basic pay-per-service plan—under which risk of equipment failure is borne by the operator—can improve performance and reduce costs.

History: Accepted by Gad Allon, operations management.

Keywords: maintenance repair • contracting • fine balance matching • service value chain • healthcare

1. Introduction

Operators of capital-intensive equipment often spend a large annual budget on maintenance so as to ensure high equipment reliability. In the medical imaging equipment industry, on which our study is based, a top-of-the-line computed tomography (CT) or magnetic resonance imaging (MRI) device costs nearly one million U.S. dollars and requires annual maintenance expenses amounting to 10% of that price (ECRI Institute 2013). Because imaging units typically last about 10 years, lifetime maintenance costs can easily approach the original equipment's purchase price. Hence, many leading manufacturers of capital equipment—including General Electric Co., Siemens AG, and Hewlett-Packard Co.—have expanded their maintenance service offerings over the last decade (Sawhney et al. 2004).

A defining feature of maintenance services is that value is coproduced (Karmarkar and Roels 2015). That is, achieving desirable outcomes (high equipment reliability and low service cost) requires operator effort in operational handling as well as service provider effort in maintenance and repair. Hence, the collaboration between the operator and service providers is critical for creating good service outcomes (Oliva and Kallenberg 2003). What governs the effectiveness of this collaboration is the maintenance service plan (MSP)—a contract that stipulates a payment structure and according to which the responsibility for equipment failure may shift between the operator and the service provider. One commonly used MSP is the basic pay-per-service contract. Under this contract, labor and parts are charged every time repair service is required, and the operator bears the bulk of equipment failure risk. Another common MSP is the full-protection contract, under which the service provider covers all maintenance costs over an agreed-upon period of time in exchange for a fixed fee. In this case, most equipment failure risk is borne by the service provider.

Our paper offers one of the first empirical analyses of how MSPs affect maintenance service outcomes. We exploit a unique data set supplied by a major medical device manufacturer to disentangle the incentive effects owing to MSPs from endogenous contract selection. This approach enables us to quantify the (relative) effect of these MSPs on failure rate, on-site visits, and remote resolutions in addition to costs of labor and replacement parts. We find that, compared to a basic contract, the full-protection contract leads to more failures and increased service costs. Because the additional costs must either be absorbed by the service provider or passed on to the operator in the form of higher fees, our results establish that the full-protection contract underperforms the basic contract at the service chain level. We demonstrate also that the service provider's greater propensity to visit operators under full-protection plans (rather than to address the issue remotely) explains the provider's contribution to the observed cost increase.

Despite the key role played by MSPs in the coproduction of maintenance services, hardly any empirical studies have explored the relative performance of these contracts. There are good reasons for this lack of empirical research to date. First, there are very few data sets of MSPs that are amenable to analysis. In addition to confidentiality issues, manufacturers have only recently begun moving into maintenance services in a significant way (Sawhney et al. 2004). Second, maintenance contracts can lead to both contract selection and contract incentive issues, and developing identification strategies that account simultaneously for these issues is challenging (Abbring et al. 2003, Chiappori and Salanié 2003).

To overcome these challenges, we use a rich data set covering the monthly service records of more than 700 medical imaging units (MRI and CT scanners) across 441 hospitals, from a major medical device manufacturer. Hospitals operate the equipment; the manufacturer provides maintenance services. The data set records the contract choices made for each unit and also their service performance and service cost, on a monthly basis, throughout the standard warranty period and during the subsequent maintenance contract period.

We exploit the structure of this unique data set to account for endogenous contract selection in our estimation of MSP incentive effects. Specifically, all hospitals have the same incentive structure during the warranty period, so any observed differences in service failures and costs among equipment over this period of time can be reasonably attributed to differences in their innate operating conditions. Adopting this account, we use a state-of-the-art matching approach to preclude any confounding effects of those innate operating conditions. Our main specification controls for potential differences in the patterns of equipment failure and service cost in the warranty period as well as for equipment type, the operator's economic importance to the service provider, time patterns, and service location. Following this procedure enables us to measure differences in service outcomes that are due to the incentive effects entailed by the type of MSP—that is, the effect on service outcomes of shifting the risk of equipment failure from one party to the other.

We find that the incentive effects arising from a full-protection plan (as compared with a basic plan) can increase the equipment failure rate by 33%. They can also increase service costs: on-site visits by 80%, service labor hours by 54%, and spare parts expended by 125%.

We also show that, in response to a reported failure, the service provider makes 55% more on-site visits to an operator under a full-protection plan than to one under a basic plan. Yet, contingent on the service provider making an on-site visit, its expenditure on labor and materials seems to be independent of contract type. In other words: the greater propensity of the service provider visiting operators under full-protection MSPs explains the increase in per-failure costs for labor and spares that is observed in the data.

Taken together, our results suggest that a full-protection plan induces the service provider to offer a higher service level (relative to a basic plan) by increasing on-site visits. It also induces the operator to rely on the service provider for maintenance and may lower the operator's own level of care for the equipment.[1] These outcomes lead to an unambiguous increase in service costs for the service chain. Given that the average participant in the U.S. health system orders one (MRI or CT) scan every three years (Smith-Bindman and Miglioretti 2012), our findings are both statistically and economically significant.

Finally, our results caution against the prevailing view that manufacturers intending to move into services should either assume more of the equipment failure risk (Oliva and Kallenberg 2003, Deloitte Research 2006) or take direct responsibility for service performance (Guajardo et al. 2012). In our setup, to the contrary, reliability deteriorates and costs increase when the service provider assumes more responsibility for equipment failure.

The paper proceeds as follows. In Section 2, we discuss the relevant literature and theory in maintenance service contracting; we then discuss the industrial setting and data in Section 3. Section 4 offers a graphical representation of the results, which leads to the development of our empirical approach in Section 5. We present our main results in Section 6 and a set of additional results and robustness checks in Section 7. Finally, in Section 8, we conclude and discuss some implications of our findings.

2. Literature Review

In the operations management (OM) literature, most theoretical research that addresses contracting focuses on the context of physical supply chains and manufacturing systems (for reviews of work in this area, see, e.g., Cachon 2003 and Nagarajan and Sošić 2008). There is comparatively less theoretical work on service contracting (Zhou and Ren 2011).

In the maintenance service literature, some papers have analyzed maintenance contracting problems within a principal–agent framework, typically where the operator is the principal and the service provider the agent. In these settings, the operator outsources maintenance work of equipment to a service provider and is assumed to have no impact on equipment failures. However, the service provider is susceptible to incentive issues because its effort in performing repairs directly affects the maintenance outcomes and yet cannot be observed by the customer. Plambeck and Zenios (2000) study such a setting with a dynamic single principal and single agent model. Kim et al. (2007) use a one-shot single principal with multiple agents to study a maintenance setting in which the service providers can affect maintenance outcomes by altering the inventory level of spare parts.

In general, however, services are coproduced in the sense that both the operator and the service provider contribute to value creation (Fuchs and Leveson 1968), and both parties may have incentive issues. For situations such as these, identifying simple contract forms that produce first-best outcomes is difficult (Bhattacharya and Lafontaine 1995, Corbett et al. 2005, Jain et al. 2013). In this stream of research, the theoretical work of Roels et al. (2010) is perhaps closest to our paper. Roels et al. compare the service chain value resulting from a basic, a full-protection, and a performance-based plan when the collaborating parties could each be susceptible to incentive distortions. Their key finding is that, depending on the context, each one of these three contract types can dominate the other two. More specifically, the performance-based plan dominates when both parties are susceptible to incentive distortions. However, the basic plan (resp., full-protection plan) dominates when only the equipment operator (resp., service provider) has incentive distortion issues. Our work provides a first empirical exploration of these theoretical predictions.

If the theoretical literature on maintenance service contracts is scarce, empirical OM work on the topic is virtually nonexistent—despite the practical importance of these issues for industry. Nonetheless, scholars have empirically studied a variety of issues in the more general context of supply chain coordination. For example, empirical papers have examined the effects of product component sharing (Ramdas and Randall 2008), information sharing (Terwiesch et al. 2005, Cui et al. 2015), and vertical integration (Novak and Stern 2008, 2009) on collaboration outcomes. These papers all focus on comparing the outcomes of collaboration and pooling of resources with the outcomes in their absence. In contrast, our paper focuses on optimizing contractual configurations conditional on collaboration.

To the best of our knowledge, Guajardo et al. (2012) is the only paper that examines empirically the performance of maintenance service contracts. The setup of our respective works differs along two dimensions. First, the contracts analyzed are of a different nature. Guajardo and colleagues focus on performance-based contracts (versus basic plans); such contracts are common in aerospace and defense industries, where most equipment owners are large entities. In contrast, we focus on full-protection plans (versus basic plans), which are much more common in the medical equipment industry and in other industries where equipment owners are small. Second, these authors focus on estimating the effects of MSPs on only one metric of service performance: the failure rate. Although we also analyze failure rates, in addition, we examine service costs (on-site visits, labor hours, spare-part costs) as well as usage rates and failure severity (the latter two on smaller sets of data).

The conclusions of the two papers are also quite different. Guajardo et al. (2012) find, for the aircraft engine industry, that performance-based contracts improve reliability over time-and-materials plans (akin to our basic plan). In other words, they find that basic plans hamper performance, which they attribute to the service provider expending less effort on repair and maintenance for operators under basic plans. In our setup, however, a basic plan improves both performance and cost (when compared with a full-protection plan). This result is driven by both the operator's and the service provider's behavior.

Finally, the economics literature includes many papers that explore the association between service contracts and performance. Similar to the OM literature, the number of empirical papers lag their theoretical counterparts; for a review of the work in this area, see Chiappori and Salanié (2003). In particular, papers on service contracts tend to focus on project-based services. Examples include, among others, IT software development (Banerjee and Duflo 2000, Gopal et al. 2003, Kalnins and Mayer 2004, Susarla et al. 2010, Susarla 2012), offshore drilling (Corts and Singh 2004), legal services (Helland 2003), construction (Bajari et al. 2014), motion picture making (Chisholm 1997), and government services (Levin and Tadelis 2010). We study instead the maintenance industry, in which the dynamics of repeated interaction introduce different types of incentives.

3. After-Sales Maintenance of Medical Equipment

Both CT scanners and MRI devices help physicians diagnose a range of conditions by providing cross-sectional views of the body's interior. Both the equipment operator (i.e., the hospital employing radiology technicians) and the service provider play a role in service outcomes. Maintenance service plans structure payments and so may affect the behavior of both parties. That is, the basic plan places a greater burden of failures on the operator, whereas the full-protection plan shifts that burden to the service provider. Shifting the burden of failure from one agent to the other

Table 1. Summary of MSP Coverage

Plan type	Labor charges	Material charges
Full-protection	Yes	Yes
Partial-protection	Yes	No
Basic	No	No

should, in theory, increase the latter's incentive to exert effort toward improving equipment reliability and reducing service costs, while reducing the former's incentives on that score. In this case, the overall effect on service outcomes depends on which of these counteracting forces prevails (Roels et al. 2010).

To study this question, we obtained contract and maintenance service records from a major medical equipment manufacturer. The data cover sales and service records of MRI and CT scanners sold by the equipment manufacturer between 2008 and 2012 in a major OECD country. The data pertain to 712 pieces of equipment, of which 57% are CT scanners.

Each new equipment sale includes a standard warranty period (usually one year); at the end of that period, the operator chooses one of the three types of MSPs. These types—in decreasing order of coverage—are (i) the full-protection plan, which offers the operator (in return for a fixed annual fee) complete protection against any labor or parts costs resulting from product failure; (ii) a partial-protection plan, in which the operator pays a fixed annual fee for all labor plus a variable amount that depends on the parts used for the repair process; and (iii) the basic plan (a.k.a. a time-and-materials contract), which requires operators to pay for all labor and parts due to repairs but with a low fixed annual fee. In addition, all MSPs cover preventive maintenance. Table 1 summarizes the key differences across these three types of plans. After the warranty period, 74% of the operators select the full-protection plan, 21% the basic plan, and 5% the intermediate, partial-protection plan. Given these figures, our analysis focuses on comparing full-protection plans with basic plans and excludes partial-protection plans.[2]

Figure 1 sketches the high-level process of how equipment failures are reported and resolved, which is standard across all operators. This process begins with an equipment operator calling in to the manufacturer's service call center to report a failure event. At this point, an electronic ticket is created to log equipment details and the problem's description. The manufacturer has a team of service call engineers, who communicate with the hospitals to understand the problem. Given remote access to the equipment for diagnosing and resolving problems, a large fraction of tickets can be resolved at this stage. For problems that cannot be so resolved, the ticket is assigned to a suitably skilled service engineer who then visits the operator's premises. After the problem is resolved, the ticket is closed, with information on the total repair hours and cost of materials logged into the system.

We track—monthly—four measures of service performance and operating costs; these measures capture all of the key process elements represented in Figure 1. The first measure is *Fails*, a count variable for the number of reported failures. The second measure, *Visits*, counts the number of on-site visits made to the operator's premises to resolve problems (some require more than one visit; others can be resolved offsite and so without any visits). The variables *Spares* and *Labor* capture the operating costs of resolving problems: *Spares* represents the cost of consumables and spare parts for maintenance and repairs, and *Labor* captures the total labor hours spent on-site.[3]

Table 2 reports summary statistics for the service performance and cost variables (observed monthly). All four variables—*Fails*, *Visits*, *Labor*, and *Spares*—are nonnegative and right-skewed, which is reflected in their large standard deviations relative to the mean. For example, on average, one piece of equipment requires a fairly low 2.8 hours of *Labor*. Yet, the standard deviation is 8.2 hours and the observed maximum is 219 hours, or 78 times the average. To tackle this issue, we use an exponential model (i.e., $E[Fails] = \exp(\beta X)$) to ensure that predictions are in the nonnegative domain, and the Poisson regression to generate estimates that are consistent in the presence of outliers and distributional misspecification (Cameron and Trivedi 2009, Wooldridge 2010).

Figure 1. High-Level Process of Reporting and Resolving Equipment Failures

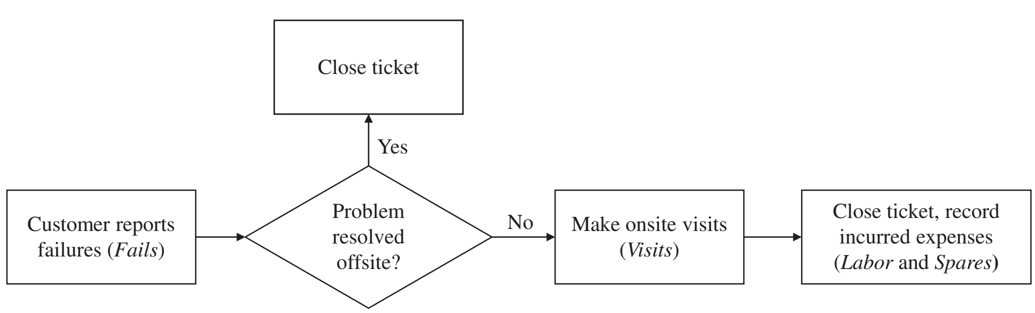


Table 2. Summary Statistics for Service Performance and Cost (712 Devices; 23,173 Monthly Observations)

			Correlation		
Variables	Description	Mean (S.D.)	1	2	3
1. *Fails*	Number of failures reported per month	1.43 (1.79)	1.00		
2. *Visits*	Number of site visits per month	1.00 (2.04)	0.70	1.00	
3. *Labor*	Hours consumed per month	2.80 (8.21)	0.54	0.85	1.00
4. *Spares*	Cost of materials consumed per month	1,502 (7,631)	0.28	0.36	0.40

Table 3. Summary Statistics for 712 Medical Scanning Devices

			Correlation	
Variables	Description	Mean (S.D.)	2	3
1. *Plan*	Basic (1), partial-protection (2), or full-protection (3)	—		
2. *Type*	MRI scanner (0) or CT scanner (1)	0.57 (0.50)	1.00	
3. *InstallDt*	Date of installation (months since January 2008)	22.7 (13.9)	−0.11	1.00
4. *CustomerSize*	Number of CT or MRI scanners purchased in the past	7.7 (11.6)	0.05	0.13
5. *Location*	Equipment's service region (1–7)	—		

Table 3 provides statistics for measures that are fixed at the equipment level. These include the plan that the operator selects (our key independent variable), the equipment type (MRI or CT scanner), the date of the equipment installation, and its physical location (one of seven regional areas covered by the service provider). Finally, we use the number of past MRI and CT scanners that the equipment operator purchased from the service provider since 1989 (the earliest date covered by our data) to create a *CustomerSize* variable that measures the operator's economic importance to the service provider.[4]

4. Visual Representation of Data and Results

Our goal is to identify the effects of the different MSPs on service performance and operating costs. Table 4 compares the averages of service outcomes across different plan types.

This table indicates that equipment tends to fail more often under full-protection than under basic plans (failure averages of 1.37 and 0.76, respectively). Equipment covered by a full-protection plan similarly consumes, on average, more repair resources: it calls for more on-site visits, more labor, and more spare parts. Table 4 does *not* indicate whether these numbers reflect self-selection effects (as when operators know their equipment has a high failure rate and therefore choose the full-protection plan) or incentive effects (as when operators and/or service providers are incentivized by the MSP to take actions that end up increasing the failure rate).

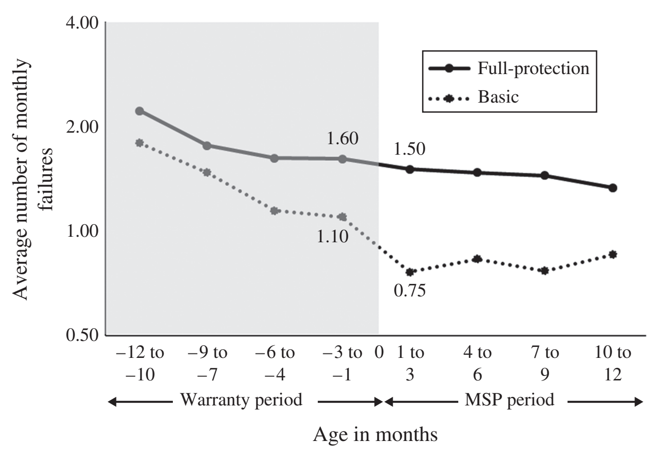
Figure 2 traces the average number of failures (on a quarterly basis) as equipment crosses the warranty period into the MSP period for the two groups of operators. Thus, the figure's solid line corresponds to operators that choose the full-protection plan and the dotted line to those that choose the basic plan. The shaded area represents the period during which equipment is still under standard warranty.

Figure 2 highlights three important aspects of the data. First, for data in the warranty period—say, from month −3 to month −1 (in the shaded area)—the average number of operator-reported failures is higher for those that selected the full-protection plan (1.60) than

Table 4. Mean Service Measures by Plan Type

Plan type	*Fails*	*Visits*	*Labor*	*Spares*
Full-protection	1.37	0.94	2.54	1,817
Basic	0.76	0.38	1.04	520

Figure 2. Average Number of Monthly Failures Over Time, Where the Warranty Expires at Time 0



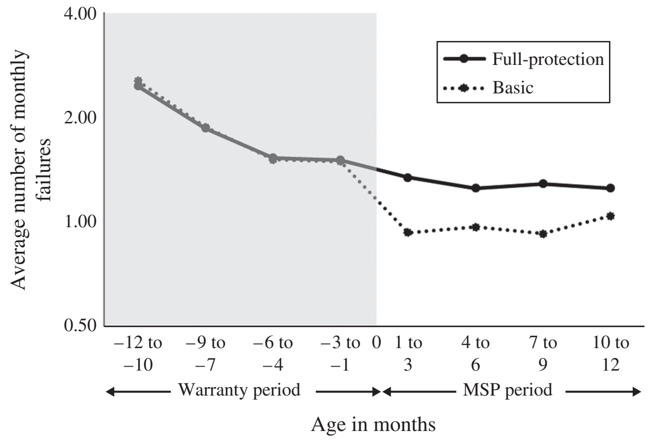
for those that selected the basic plan (1.10). The difference (of 0.50 failures per month) between the two groups of operators during this period provides visual evidence of contract selection; that is, operators do not choose MSPs randomly but instead choose MSPs with more protection if they have higher failure rates.

Second, Figure 2's dotted line, which marks the equipment failure rates for customers on the basic plan, exhibits a sharp drop in reported failures as the operators transition from the warranty period into the MSP period (from 1.10 failures in months −3 to −1 down to 0.75 failures in months 1 to 3). This drop is less pronounced for customers that select the full-protection plan (1.60 versus 1.50). These concurrent changes result in a widening of the gap—as customers transition from the warranty to the MSP period—between the line representing customers that select the full-protection plan and the one representing those that select the basic plan. That widening of the gap (e.g., from $1.60 - 1.10 = 0.50$ to $1.50 - 0.75 = 0.75$) reflects the impact of MSPs on service performance that is due to the service plans' different incentive structures.

The third important aspect of our data is that the number of failures declines during the warranty period, and this decline appears faster for customers that select the basic plan. This declining failure rate is consistent with the so-called bathtub curve, whereby failures typically first decrease, then flatten out, and finally increase with equipment age (Nowlan and Heap 1978, Smith and Oren 1980, Aarset 1987, Block and Savits 1997); an early decrease in the failure rate is usually attributable to the detection and resolution of (nonrepeating) defects in manufacturing. The difference in trends further suggests that an equipment under a basic plan may be located on a different part of this curve than is an equipment under a full protection plan. This possibility constitutes one of our main estimation challenges, since that difference (in location on the bathtub curve) might itself at least partly explain the widening gap observed when the warranty expires.

That said, the data in the warranty period allows us to observe innate differences across equipment that we leverage to identify incentive effects. Specifically, to tackle the issues of contract selection and failure trends, we match equipment with similar failure trends during the warranty period (but that ended up on different plans), thereby eliminating systematic differences during the warranty period across the two groups (details are presented in the next section). Therefore, any remaining difference after the warranty ends must capture the MSP effect. Figure 3 presents the equivalent of Figure 2 *after the matching procedure*. Here, we see that, during the warranty period, the trends are now essentially identical. Even so, there is once again a sharp divergence in reported failures almost as soon as the warranty period ends and MSP incentives come into play. In Section 6, we establish that this incentive effect is statistically significant.

Figure 3. Average Number of Monthly Failures Over Time for the Matched Population Where the Warranty Expires at Time 0



Finally, matching is a natural approach in our setup. While it is challenging to control for the differences in trends and curvatures of equipment in a regression framework, matching allows us to do this nonparametrically. Identification using the entire sample in standard regressions would heavily rely on extrapolation (i.e., comparisons using innately dissimilar units). By contrast, identification under matching relies on the two pools of equipment where we observe significant overlap (as in Figure 3).

5. Estimation Approach

Our approach consists of three stages. In the first, we rely on data from the warranty period to estimate parameters for equipment failure patterns. In the second stage, we match equipment in the full-protection plan against equipment in the basic plan in terms of those estimated failure patterns and other equipment characteristics. Finally, we use observations from the post-warranty period to estimate the incentive effect.

5.1. Estimating Parameters of Failure Patterns in the Warranty Period

The main confounds with respect to our data arise because each device may have their own levels and trends of failures. We use a three-parameter $(\alpha_i, \beta_i, \theta_i)$ model to capture the failure patterns of each piece of equipment i (in the warranty period) conditional on that unit's age in months:

$$E[Fails \mid i, Age] = \exp\{\alpha_i + \beta_i(Age) + \theta_i(Age^2)\}. \quad (1)$$

Note that the right-hand side (RHS) of Equation (1) is enclosed in an exponent so that predictions will remain in the nonnegative domain. Parameters α_i, β_i,

Table 5. Distribution of $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\theta}$

			Quantiles (%)				
Estimates	Mean	S.D.	5	25	50	75	95
$\hat{\alpha}$	0.119	1.077	−1.80	−0.24	0.34	0.80	1.33
$\hat{\beta}$	−0.062	0.269	−0.31	−0.11	−0.03	0.03	0.12
$\hat{\theta}$	−0.005	0.047	−0.06	−0.02	0.00	0.02	0.05

and θ_i denote, respectively, the level, trend, and curvature of equipment i's failures. Equation (1) can be estimated separately for each equipment, or equivalently in a joint estimation using a single Poisson regression (with level, trend, and curvature identifiers for each equipment).

In essence, we aggregate the failure patterns of the warranty period into three parameters. Matching directly on the data runs the risk of matching on idiosyncratic month-to-month failure patterns. Consistent with the literature (Nowlan and Heap 1978, Smith and Oren 1980, Aarset 1987, Block and Savits 1997), we therefore assume that an underlying failure curve generates the data (together with other sources of random variations generating the final monthly failures). Additionally, our quadratic model is sufficiently general to capture all six patterns of failure curves identified by Nowlan and Heap (1978), including the bathtub curve. We note that the generality is warranted in this case: the model fit for a quadratic model is significantly better than the model fit for a linear model.[5]

Table 5 reports the distributional statistics for the estimated parameters.[6] The magnitudes of the estimation errors (the standard errors of $\hat{\alpha}$ of 0.44, $\hat{\beta}$ of 0.12, and $\hat{\theta}$ of 0.03) are small compared to the variation of the failure curve parameters across equipment reported in the table (standard deviation of $\hat{\alpha}$ of 1.08, $\hat{\beta}$ of 0.27, and $\hat{\theta}$ of 0.05). Thus, between-equipment variation accounts for most of the total variation (86% for α, 84% for β, and 78% for θ),[7] and the variation in our data is mainly driven by cross-equipment variation rather than estimation errors. As a result, the imprecision of the failure curve estimates should not significantly impact the precision of the matching procedure we present next. We check this further in Section 7 with a bootstrap approach that directly accounts for the uncertainty in the failure curve estimates. This alternative approach produces very similar final estimates, and all of our results remain statistically significant.

5.2. Matching

We follow an optimal matching approach with "fine balance" (Zubizarreta 2012, p. 1360) to create two populations of equipment: one each for the basic and full protection plans. This approach matches exactly on a one-to-one basis the variables that have a significant impact on identification (such as the failure curve). Less critical variables are matched on a population level—i.e., the matching is done so that the two populations are distributionally similar. The flexibility of the method allows us to achieve a close match between equipment without significant loss of data (Rosenbaum et al. 2007, Yang et al. 2012).[8]

5.2.1. Variables for Matching. To ensure that the two populations of equipment are similar, we match them on 10 variables as follows. First and foremost, we match the three estimated failure parameters ($\hat{\alpha}$, $\hat{\beta}$, and $\hat{\theta}$) to ensure that there are no systematic differences in equipment—with respect to their level of failure or position on the bathtub curve—during the warranty period.

Second, we ensure the absence of any aggregate differences in service response by matching also on three service cost measures: the average monthly number of site visits made on account of each device during the warranty period (*AvVisits*), the average hours of labor spent (*AvLabor*), and the average cost expended on spare parts (*AvSpares*).

Finally, we match on the equipment's type, customer size, installation date, and location (variables that are fixed at the equipment level; see Table 3). Matching on equipment type controls for potential systematic differences (regarding, e.g., usage or failure patterns) between MRI and CT scanners. Matching on customer size ensures that we account for the equipment operator's economic importance to the service provider. Matching on installation date (i.e., months elapsed since January 2008) ensures that our observations begin and end at similar times; in this way, we control for general time trends.[9] Finally, matching on location ensures that geographic differences across service regions (e.g., distance from the service provider) do not affect our estimates.

5.2.2. Optimal Matching with Fine Balance. We formulate the matching problem as a mixed-integer program whose objective is to maximize the number of paired samples subject to balance constraints (Zubizarreta 2012, Zubizarreta et al. 2014). For the most critical features—and in particular the failure curve—these constraints reflect an exact match of equipment. For features that are less critical, the constraints enforce a *fine balance* (Rosenbaum et al. 2007, Yang et al. 2012): instead of matching such features exactly, the constraints guarantee that the *distribution* of equipment with respect to the considered features is similar across the two populations. Following this approach allows us to limit data loss while retaining accurate matches.

We start by identifying the location of 1s in a matching matrix $\mathbf{M}$ with elements M_{bf}, where b indexes equipment in $\mathbf{B}$ (i.e., on the basic plan) and f indexes equipment in $\mathbf{F}$ (on the full-protection plan). Note that $M_{bf} = 1$ only if equipment b is matched to equipment f

(and otherwise equals 0). The objective of maximizing the number of matches is formally stated as follows:

$$\max_{M_{bf}\in\{0,1\},\, b\in B,\, f\in F} \sum_{b\in B,\, f\in F} M_{bf}. \tag{2}$$

We now introduce a series of constraints that represent the matching requirements. First, we implement a one-to-one match; that is, a scanner on the basic plan must be paired with exactly one scanner on the full-protection plan. Formally, we have

$$\begin{aligned} &\forall f\in\mathbf{F}, \quad \textstyle\sum_{b\in B} M_{bf}\le 1,\\ &\forall b\in\mathbf{B}, \quad \textstyle\sum_{f\in F} M_{bf}\le 1. \end{aligned} \tag{3}$$

Next, we match on equipment features. We match exactly on the equipment type and failure curve. So if $M_{bf}=1$, then we need that $Type_b = Type_f$ and that the parameters $\hat\alpha_b$, $\hat\beta_b$, and $\hat\theta_b$ be, respectively, within ε_α, ε_β, and ε_θ of the parameters $\hat\alpha_f$, $\hat\beta_f$, and $\hat\theta_f$. We choose ε such that the distance between the parameter values is less than 0.37 of the interquartile range.[10] For instance, $\varepsilon_\alpha = 0.37\times(0.80-(-0.24)) = 0.38$ from the 75th and 25th quartile in Table 5.

The result is the following set of constraints. For all $b\in\mathbf{B}$ and $f\in\mathbf{F}$:

$$\begin{aligned} &M_{bf}(Type_b - Type_f) = 0;\\ &M_{bf}|\alpha_b-\alpha_f|\le\varepsilon_\alpha,\\ &M_{bf}|\beta_b-\beta_f|\le\varepsilon_\beta,\\ &M_{bf}|\theta_b-\theta_f|\le\varepsilon_\theta. \end{aligned} \tag{4}$$

We match the remaining features via the fine-balance approach. Consider the categorical variable *Location*, which locates each piece of equipment i into one of seven distinct service regions. Our constraint must guarantee that the distribution of equipment over locations is the same across the two matched populations. Denoting by $I(\cdot)$ the indicator function, we have the following constraints:

$$\forall j\in\{1,\dots,7\}, \quad \sum_{f\in F,\, b\in B} M_{bf} I(Location_b = j) = \sum_{f\in F,\, b\in B} M_{bf} I(Location_f = j). \tag{5}$$

The equivalent fine-balance matching for continuous variables consists of matching on quantiles (Zubizarreta 2012), so we constrain the two distributions to have the same quartiles (median, 25th percentile, and 75th percentile). We employ this technique to match five different variables: (i) the average monthly number of site visits made to each piece of equipment during the warranty period (*AvVisits*), (ii) the average hours of labor spent (*AvLabor*), (iii) the average cost of spare parts (*AvSpares*)[11], (iv) *CustomerSize*, and (v) the calendar month during which the equipment was installed (*InstallDt*). We denote by F_v^{-1} the inverse cumulative distribution of variable v; hence, $I(v_b \ge F_v^{-1}(p)) = 1$ if and only if the variable v for equipment b is at no less than the pth percentile of the variable's distribution. Thus, we have the following constraints:

$$\begin{aligned} &\forall v\in\{AvVisits, AvLabor, AvSpares, CustomerSize,\\ &\qquad InstallDt\},\\ &\forall p\in\{0.25, 0.5, 0.75\},\\ &\sum_{f\in F,\, b\in B} M_{bf} I(v_b\ge F_v^{-1}(p))\\ &\quad = \sum_{f\in F,\, b\in B} M_{bf} I(v_f\ge F_v^{-1}(p)). \end{aligned} \tag{6}$$

In short, the matching problem amounts to maximizing objective (2) subject to constraints (3)–(6). Thus, our approach maximizes the number of paired samples subject to fine, one-to-one constraints on equipment type and the failure curve and subject to distributional constraints on service cost measures, geographical location, customer size, and installation date.[12]

5.2.3. Outcome of Matching.

Using the matching approach just described, we identify a total of 188 units (94 each under the basic and full-protection plans). Table 6 compares the statistics of our variables across the two populations. We apply three tests to check the quality of the match. Two of these tests are discussed here; the more extensive "placebo" test is presented in Section 7.1.

The first is a univariate test that checks, for each variable, whether the difference of the means is less than one-fifth of a standard deviation (Cochran 1968, Cohen 1988, Imai et al. 2008). When that is the case, group

Table 6. Comparison of Univariate Statistics After Matching (188 Scanning Devices)

	Basic	Full-protection	Normalized difference
Variables	μ_B (σ_B)	μ_F (σ_F)	$(\mu_F-\mu_B)/\sigma_{\text{pooled}}$
Type	0.50 (0.50)	0.50 (0.50)	0.00
$\hat\alpha$	0.27 (0.81)	0.29 (0.76)	0.01
$\hat\beta$	−0.05 (0.09)	−0.05 (0.10)	−0.01
$\hat\theta$	0.00 (0.02)	0.00 (0.02)	0.00
AvVisits	1.30 (1.07)	1.40 (1.09)	0.10
AvLabor	3.74 (3.40)	4.01 (3.68)	0.08
AvSpares	808 (926)	765 (907)	−0.04
CustomerSize	8.77 (12.3)	8.63 (13.2)	−0.01
InstallDt	22.5 (11.0)	23.1 (10.3)	0.06
Location	—	—	—

Notes. μ_B and σ_B (resp., μ_F and σ_F) are the mean and standard deviation for the basic (resp., full) plan. The divisor σ_{pooled} is the pooled standard deviation, which is calculated as $\sigma_{\text{pooled}} = \sqrt{0.5(\sigma_B^2+\sigma_F^2)}$. Statistics are not shown for *Location*, which is both categorical and perfectly balanced.

membership (in either a basic or a full-protection plan) explains less than 1% of each variable's total variance. We find that all variables satisfy this strict criterion—i.e., the magnitude of the numbers in the last column of Table 6 should be less than 1/5.[13] The second test checks for whether the variables can jointly predict which contract an operator chooses after the warranty period (Imbens 2004). Indeed, if our matches are of good quality, then equipment characteristics should have no predictive power with regard to contract choice. We use a logit regression in which the *Plan* dummy variable is set to 1 if the equipment operator chose the full-protection plan (and to 0 otherwise). We run the regression specified in 7 below and find that none of the coefficients is statistically significant. Of critical importance is that, when we compare this model with a simpler model that involves just a constant (i.e., $\text{logit}P[I(Plan_i = \text{full-protection})] = c$), we find that the model specified in Equation (7) does not fit any better than the simple model: a likelihood ratio test compares the likelihood of both models and finds no statistical significance ($\chi^2 = 2.29$, $p = 0.999$).

$$\begin{aligned}\text{logit}P[I(Plan_i = \text{full-protection})] &= c_\alpha \alpha_i + c_\beta \beta_i + c_\theta \theta_i + c_v AvVisits + c_l AvLabor \\ &\quad + c_s AvSpares + c_c CustomerSize + c_d InstallDt \\ &\quad + \sum_{j\in\{1,\ldots,7\}} c_j I(Location_i = j). \qquad (7)\end{aligned}$$

Figure 3 illustrates the overall quality of the match and shows that paired equipment in the full-protection and the basic plan exhibit strongly similar failure rate curves during the warranty period.

5.3. Estimating Incentive Effects

We now turn to the estimation of incentive effects, which are captured by the difference in service outcomes during the postwarranty period. (For service failures, this corresponds to the gap observed in Figure 3 and discussed in Section 4.) Equation (8) specifies the regression setup:

$$\begin{aligned}&E[\text{Service outcome} \mid \text{MSP}] \\ &\quad = \exp\{\delta I(\text{MSP} = \text{full-protection}) + c\}. \qquad (8)\end{aligned}$$

Here, "Service outcome" may correspond to the number of failures, site visits, labor hours, or spares consumed. We are most interested in δ because it represents the incentive effect arising from shifting an operator from the basic plan to a full-protection plan. The constant c models the baseline aggregate service outcome—that is, when equipment is on the basic plan. We do not incorporate any control variables since we account for them in the matching process (Imbens 2004).[14]

Given that the service outcome measures are nonnegative and right-skewed, we first enclose the RHS in an exponent to ensure that predictions are in the nonnegative domain. We also use the Poisson regression to estimate the model (which, as mentioned in Section 3, is unbiased in the presence of outliers and distributional misspecification). All reported standard errors are clustered by the equipment, so they are robust to the potential presence of autocorrelation and heteroskedasticity.

6. Results

6.1. Effect of MSP on Service Outcomes

Table 7 presents results for the regressions. Consider first our model in which the dependent variable (DV) is *Fails*. The analysis reveals a statistically significant ($p < 0.05$) and positive effect of full coverage—as compared with basic coverage—on the number of reported failures. The mean estimate of the effect δ is 0.29, which corresponds to a 33% increase in the failure rate.[15]

Consider now the model with dependent variable *Visits*. The effect of a full-protection plan on the number of site visits made by the service provider is again statistically significant ($p < 0.001$). Moreover, for a 33% increase in failure rate, the service provider increases the number of on-site visits by 80%. These results suggest that, given a reported failure, the service provider is more likely to visit an operator under *Full* coverage. (This hypothesis is tested formally in Section 6.2.)

Similarly, the coefficients δ in the *Labor* and *Spares* models indicate statistically significant increases: in *Labor* ($p < 0.05$) and also in *Spares* ($p < 0.01$). Specifically, *Labor* increases by 54% while *Spares* increases by 125%.

Altogether, these findings establish that MSPs have a large and statistically significant incentive distortion effect, leading to large increases in costs for servicing equipment under the full-protection plan as compared with the basic plan. It is certainly possible that both parties contribute to this cost increase; yet, in theory, the equipment operator's contribution should be more significant because that operator pays a fixed annual cost regardless of the final repair cost under the full-protection plan.

Table 7. Poisson Regression Estimating the Effect of MSP on Service Performance and Operational Costs

DV:	*Fails*	*Visits*	*Labor*	*Spares*
δ	0.29 (0.12)*	0.59 (0.15)***	0.43 (0.21)*	0.81 (0.30)**
Observations	2,085	2,085	2,085	2,085
Log-likelihood	−3,257	−2,717	−8,106	−1,828,199

Note. Standard errors are clustered by equipment.
$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

6.2. Effect of MSP on Service Provider

To explore this incentive distortion effect further, we now examine in greater detail how the service provider responds to failures. For example, the service provider's propensity to make an on-site visit (in response to a reported failure) may vary depending on the equipment operator's type of MSP. The service provider could also, in theory, vary its repair approach—for example, by spending more time on the careful diagnosis of difficult problems or by deciding to replace not only a failed component but rather all interrelated components. Such incentive effects would be captured by changes in on-site visit rates (i.e., number of visits per reported failure) and in the hours and costs expended per visit.

To explore these possibilities, we create the variables *Visits/Fail*, *Labor/Visit*, and *Spares/Visit*—dropping months during which we observed no failures or (for the latter two variables) no visits. We analyze these dependent variables using the same setup as in Equation (8), and our results are reported in Table 8.

The coefficient δ for *Visits/Fail* is statistically significant ($p < 0.001$). Its value of 0.44 indicates a 55% increase in the number of visits per failure. This result is *not* consistent with the notion that a service provider would save on repairs when it bears all of their costs. In fact, when the service provider bears these costs, it tends to make *more* on-site visits.

Finally, the coefficients derived in the regressions on *Labor/Visit* and *Spares/Visit* are not statistically significant. So, conditional on the service provider dispatching an engineer, there are no statistically significant differences in how service engineers repair equipment due to the MSP chosen by an operator. One possible explanation for this is that the frequency of site visits is an indication of responsiveness, which allows the service provider to further signal the added value of the full-protection MSP; but, once on-site, the engineer seems to follow standard procedures.

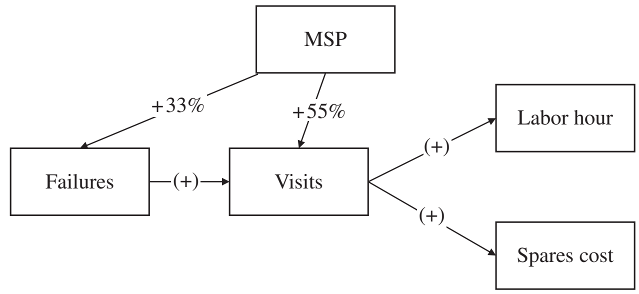
Figure 4 summarizes our findings. The diagram's arrows indicate that the full-protection MSP has two types of effects. First, it drives a 33% increase in reporting of failures by the operator; second, it induces the service provider to make 55% more on-site visits in response to reported failures. When combined, these two effects account for the large increases in labor costs and spare-parts cost (54% and 125%, resp.) described in the main analysis.

Table 8. Poisson Regression Estimating the Effect of MSP on Failure Response

DV:	*Visits/Fail*	*Labor/Visit*	*Spares/Visit*
δ	0.44 (0.07)***	−0.06 (0.10)	0.19 (0.19)
Observations	1,143	672	672
Log-likelihood	−1,052	−1,428	−386,899

Note. Standard errors are clustered by equipment.
***$p < 0.001$.

Figure 4. Effect of Full-Protection MSP on Service Outcomes



7. Robustness and Additional Analysis

7.1. Placebo Tests

We test the validity of our approach by performing a "placebo test" (Angrist and Pischke 2008). The idea is to use data from the warranty period only, then to break that period into two subperiods, and finally to repeat our analysis while supposing that the warranty period had ended (and the MSP period had begun) on completion of the first subperiod. If there were selection factors that our main model did not adequately account for, then the results of these placebo tests should show some statistically significant effect.

We break the standard warranty period into two subperiods as follows: the first subperiod covers all *except* the last three months of the warranty period, and the second subperiod covers those remaining three months. Table 9 depicts the results of this analysis.

Table 10 shows the result of an alternative setup where the first subperiod covers all except the last six months. We can see that in none of the models are any of the δ coefficients statistically significant at the 5% level. Indeed, the mean effect is much smaller than (and

Table 9. Poisson Regression Model Assuming Warranty Period Ends Three Months Earlier

DV:	*Fails*	*Visits*	*Labor*	*Spares*
δ	0.04 (0.13)	0.22 (0.24)	0.05 (0.32)	−0.26 (0.39)
Observations	342	342	342	342
Log-likelihood	−595	−552	−1,591	−394,618

Note. Standard errors are clustered by equipment.

Table 10. Poisson Regression Model Assuming Warranty Period Ends Six Months Earlier

DV:	*Fails*	*Visits*	*Labor*	*Spares*
δ	0.10 (0.19)	0.14 (0.30)	−0.01 (0.42)	0.01 (0.48)
Observations	276	276	276	276
Log-likelihood	−496	−469	−1,339	−336,961

Note. Standard errors are clustered by equipment.

Table 11. Comparison of Poisson Regression Estimates Based on Our Main Model vs. Bootstrapping Failure Curves

DV:	*Fails*	*Visits*	*Labor*	*Spares*
Main model	0.29 (0.12)*	0.59 (0.15)***	0.43 (0.21)*	0.81 (0.30)**
Bootstrap model	0.30 (0.12)*	0.60 (0.16)***	0.46 (0.23)*	0.87 (0.32)**

Note. Standard errors clustered on equipment.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

directionally inconsistent with) the effects obtained in our main analysis.

In other words, the two groups that are matched up to the end of the first subperiod continue to have parallel trends from that time onward until the warranty period ends. This finding confirms that our observed effects in the postwarranty period stem from the contract type and are not due to inadequately controlling for selection factors (e.g., the bathtub curve).

7.2. Errors Arising from Matching on Estimates

To further understand the robustness of our findings to the errors in the failure curve parameter estimates, we compare our final estimates with the ones obtained by bootstrapping. Using the joint distribution of $\hat{\alpha}$, $\hat{\beta}$, $\hat{\theta}$ established in Section 5.1, we first create 500 bootstrapped samples (each sample consists of 465 sets of failure curve parameter estimates; i.e., one set of $\hat{\alpha}$, $\hat{\beta}$, $\hat{\theta}$ for each equipment). We then proceed with matching and estimating the incentive effects based on these bootstrapped samples. Table 11 compares the results using bootstrapping against our main model. As shown in the table, the bootstrapping method produces estimates that are very close to the original ones, and all effects remain statistically significant.

7.3. Interactions

In this section, we extend the main analysis by considering possible interactions. More specifically, we test whether our results differ across customer size, equipment type, or time. To limit the model's complexity, we test for these effects separately. In none of these cases do we detect any significant interaction effects.

Customer size may interact with contractual incentives if the service provider offers special treatment to only those operators that are large *and* on the full-protection plan. We can test for this possibility by interacting δ with *CustomerSize*, thereby creating the variable $\delta \times CustomerSize$, and then including this term—along with *CustomerSize*—in the main regression.[16] Table 12 reports the results from this estimation. Note that the coefficient for $\delta \times CustomerSize$ is not significant, indicating that the incentive effect does not seem to increase with larger customers.

In the same vein, we also test for differences in the incentive effects across scanner type: MRI versus CT. Here, we interact $\delta \times I(Type = \text{MRI})$ and include the term $I(Type = \text{MRI})$. Table 13 reports the results. We see that, as compared with CT scanners, MRI scanners (on average) have higher failure rates and incur greater service costs. Nonetheless, the coefficients for the interaction terms are again not statistically significant.

Finally, there may be variation in effects across time if, for example, the operator and service provider learn how to cooperate better. We implement a test of this possibility by interacting δ with *Age* (i.e., $\delta \times Age$) and then including this term, together with *Age*, in the regression.[17] Results are given in Table 14. We see no evidence—over a one-year time frame—of any changes in incentives.

Table 12. Poisson Regression Model Including Interaction with Customer Size

DV:	*Fails*	*Visits*	*Labor*	*Spares*
δ	0.28 (0.11)*	0.59 (0.15)***	0.41 (0.21)*	0.83 (0.30)**
CustomerSize	0.00 (0.01)	0.00 (0.01)	−0.00 (0.01)	−0.02 (0.01)
$\delta \times CustomerSize$	−0.00 (0.01)	−0.01 (0.01)	−0.01 (0.01)	0.01 (0.02)
Observations	2,085	2,085	2,085	2,085
Log-likelihood	−3,255	−2,714	−8,073	−1,816,573

Note. Standard errors are clustered by equipment.
$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Table 13. Poisson Regression Model Including Interaction with Equipment Type

DV:	*Fails*	*Visits*	*Labor*	*Spares*
δ	0.48 (0.19)*	0.80 (0.23)***	0.69 (0.20)*	1.41 (0.37)***
$I(Type = \text{MRI})$	0.81 (0.18)***	0.73 (0.24)**	0.70 (0.34)*	1.18 (0.42)**
$\delta \times I(Type = \text{MRI})$	−0.34 (0.22)	−0.37 (0.29)	−0.46 (0.40)	−0.97 (0.55)
Observations	2,085	2,085	2,085	2,085
Log-likelihood	−3,149	−2,674	−8,012	−1,797,356

Note. Standard errors are clustered by equipment.
$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

Table 14. Poisson Regression Model Including Interaction with Age

DV:	*Fails*	*Visits*	*Labor*	*Spares*
δ	0.29 (0.12)*	0.59 (0.15)***	0.48 (0.20)*	0.81 (0.29)**
Age	−0.00 (0.01)	−0.01 (0.03)	−0.01 (0.01)	−0.02 (0.05)
$\delta \times Age$	−0.01 (0.01)	0.01 (0.03)	0.05 (0.05)	0.02 (0.07)
Observations	2,085	2,085	2,085	2,085
Log-likelihood	−3,257	−2,717	−8,057	−1,827,697

Note. Standard errors are clustered by equipment.
$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

8. Discussion and Conclusion

We have established that—in the context of medical equipment maintenance—a fixed-fee, full-protection plan leads to more failures and higher service costs than does a basic, time-and-materials plan. For this reason, service providers should be wary of assuming more responsibility for equipment failure outcomes (as they do when absolving operators under the full-protection plan of repair costs). Indeed, Roels et al. (2010) have theorized that this may be a possible scenario, and our work provides the first supporting empirical evidence.

The analysis presented here also establishes that both the equipment operator and the service provider contribute to increased service costs. Operators on full-protection plans apparently reduce the level of their own care of the equipment.[18] In contrast, the service provider expends more resources—in terms of a higher on-site response rate—to equipment under a full-protection plan.[19]

It is noteworthy that, according to the documents available to us, neither the contract terms nor the maintenance process itself dictates that more on-site visits (or any other special treatment) be given to operators under a full-protection plan. Furthermore, account managers—who could have an incentive to affect the quality of service afforded certain operators—are not actually involved in the maintenance process. The increase in on-site visits may therefore result from the intrinsic motivation of the service provider and/or broader organization-level culture (see, e.g., Kreps 1997). This, in turn, leads to additional and perhaps unnecessary costs.

Thus, our study documents the presence of significant frictions that result in a decrease in service chain value relative to the ideal, first-best outcome. In other words, a centralized decision maker who managed both the service provider and the equipment operator could generate higher profits (by reducing equipment maintenance costs) than the sum of their currently achieved individual profits. In this sense, our study reveals that there are costly inefficiencies in the medical equipment industry and that these inefficiencies are significant—especially considering the size of the industry.

This research makes clear the need for better coordination mechanisms between service provider and equipment operator. It may be that alternative contracts could result in Pareto improvements over the outcomes of existing contracts. Coordination might also be improved by alternative business models in which, for instance, the service provider directly operates the equipment at the hospital.

One limitation of our work is the absence of data on preventive maintenance. That being said, the "reactive" maintenance events we study constitute most of all interventions (preventive *and* reactive) in our setup. We therefore do not expect that accounting for preventive maintenance activities would significantly alter our findings. Yet, if such an effect does exist, then in theory it should be in the same direction as our main result. Hence, the increase in cost and failures observed in the data would be even greater.

Finally, we are limited by the lack of revenue data (owing to commercial sensitivity) and are prevented from discussing the implications of our results for either the service provider's or the equipment operator's profits. An interesting avenue for future research, therefore, is to determine if the loss of value we characterize in this paper is actually priced in or if the service provider absorbs the cost.

Acknowledgments

The authors thank department editor Gad Allon, the associate editor, and two referees for their constructive feedback. They also thank their industry partner (kept anonymous) for helpful discussions, insights, and input.

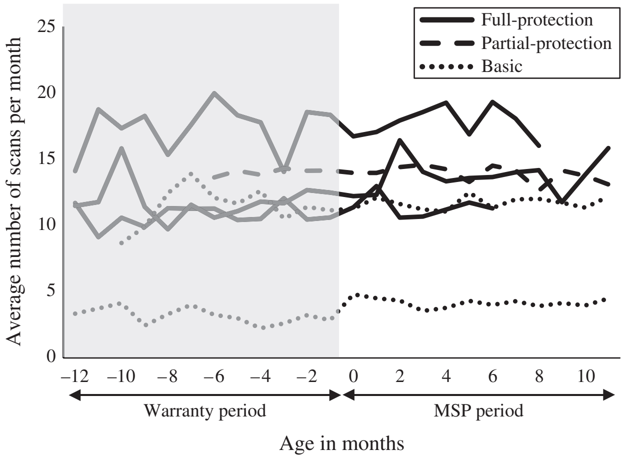
Appendix. Analyses with Supplementary Data

In this appendix, we detail analyses that use supplementary data sets. We use these data sets to show that contracts have negligible effects on equipment usage and failure severity.

A.1. Equipment Use

Here, we use scanning data from a small sample of 471 observed months (covering 22 pieces of equipment) to investigate directly whether plans affect usage. The use data (measured as number of scans per month, but masked by a multiplication factor) were made available through a trial program whereby the equipment's status is tracked remotely in real time. Figure A.1 plots use data over time for a sample of six pieces of equipment.

Figure A.1. Number of Scans Across Time



Observe that the number of scans differs considerably from one equipment operator to the next. For a given operator, however, the number of scans is quite stable over time. More crucially, the postwarranty change in contract type does *not* result in visually detectable changes in the number of scans.

We use fixed effects in formally testing the relation between MSP type and equipment usage. As compared with the matching approach deployed in the main text, a fixed-effects approach differs in two ways. First, it leverages all of the data for inference (and hence is more "data efficient" for small data sets). Second, whereas matching removes any preexisting differences with regard to the warranty period, inference based on fixed effects relies on difference in differences (i.e., whether the aggregate gap in usage between those under a full-protection plan versus a basic plan *changes* after the warranty period; see Angrist and Pischke 2008). The specification is given in Equation (A.1), where (as in our main specification) δ captures the incentive effects of the full-protection plan (relative to the basic plan). Once again, we use a Poisson regression to estimate the model:

$$\begin{aligned} &E[Scans \mid \text{Equipment } i, \text{Month } m] \\ &\quad = \exp\{c_i + \delta_w I(Contract_{im} = \text{warranty}) \\ &\qquad + \delta I(Contract_{im} = \text{full-protection})\}. \end{aligned} \tag{A.1}$$

The results are reported in Table A.1; note that the δ coefficient is not statistically significant. So in line with Ning et al. (2018), who study incentive effects on usage in pay-per-print services, we find that contracts have a negligible incentive effect on equipment usage.

Table A.1. Effect of Plan on Use (Poisson with Fixed Effects)

DV:	*Scans*
δ_w (warranty)	−0.09 (0.05)
δ	0.01 (0.07)
Equipment fixed effects	Yes
Observations	441
Log-likelihood	−764

Note. Standard errors are clustered by equipment.

Table A.2. Effect of MSP on Failure Severity (Conditional Logit)

DV:	*Severity*
δ_w (warranty)	0.01 (0.43)
δ	0.06 (0.48)
Equipment fixed effects	Yes
Observations	3,678
Log-likelihood	−1,649

Note. Standard errors are clustered by equipment.

A.2. Failure Severity

Our main analysis focuses on failure counts, but failures are heterogeneous, and the incentive effects we have uncovered could similarly affect failure severity. Here, we test if there are variations in failure severity due to MSP changes. This provides a direct test of whether the increase in on-site visits is driven by variations in failure severity. It also indirectly tests whether contracts affect reporting behavior.

Given the very nature of reporting behavior, direct data on it will (almost) never be available. Nonetheless, failure severity is a natural indirect measure. Indeed, severe failures correspond to those that shut down a piece of equipment because they cannot be fixed by the operator; thus, they can hardly go unreported. In contrast, the operator may be able to fix a nonsevere failure or at least to continue using the equipment despite the problem. It follows that any changes in the level of reporting would be detectable through changes in the mix of severity of reported failures.

To help disentangle this issue, we obtained data covering the severity of each failure over a six-month period (about 4,000 failure occurrences). Failures can either involve a loss of functionality (e.g., equipment won't power up, cooler not working, table not feeding; coded by the service provider as 1) or involve only minor issues (abnormal noises, minor software issues, etc.; coded as 0). Nearly a quarter of all failure events are severe.

We test the relationship between failure severity and reporting tendencies by using a fixed-effects setup similar to that used in Section A.1 (but with conditional logit since the dependent variable now takes binary values). The results of this analysis are reported in Table A.2. Observe that, once again, the δ coefficient is not statistically significant. Thus, we find no statistical evidence that MSPs have an effect on the severity of equipment failures.

Endnotes

[1] Other possibilities explaining the increase in failure rate is an increase in usage rate, or an increased propensity to report failures. Our analysis using supplementary data sets did not find support for these two alternative explanations (see the appendix).

[2] The number of scanners under partial protection is too small to allow reliable statistical statements. Hence, our main analysis addresses the incentive effects of the basic and full-protection plans only.

[3] We scaled both *Labor* and *Spares* with a fixed (positive) multiplicative factor to mask the absolute magnitude of the variables. This transformation does not affect our estimation because we use an exponential setup, which reports results as percentage changes from a given base.

[4] *CustomerSize* does not change frequently over time, and we measure this variable at the installation date.

[5] The likelihood ratio test produces $LR = 837$, $P(LR > \chi^2) < 0.0001$.

[6] The distribution *excludes* all equipment on partial-protection plans and includes only equipment with a warranty period of exactly 13 months (i.e., the modal length), for a total of 465 medical scanning devices.

[7] Using the formula $\sigma_B^2/(\sigma_B^2 + \sigma_W^2)$, where σ_B is between equipment standard deviation, and σ_W is within standard deviation arising from estimation errors.

[8] To further appreciate this point, consider an alternative close-distance matching approach that requires identifying pairs of devices that are *close across all covariates*. While theoretically ideal, identifying pairs such as these is difficult with a large number of matching variables, and the approach results in very small and sensitive samples (Stuart 2010). One solution for this issue is to relax how close pairs need to be as in a "coarsened exact matching" (CEM) approach (Iacus et al. 2011). While using CEM leads to qualitatively similar results (not reported here), the approach produces a sample that is smaller with matches on the failure curves that are less precise.

[9] The effect of this procedure is similar to that of controlling for calendar year-month fixed effects in regressions. We could instead control for such fixed effects later in the regression, but it is preferable to control for all key variables up front in the matching process (Imbens 2004).

[10] 0.37 of interquartile range (IQR) is equivalent to 0.5 standard deviation for a normally distributed variable. We use IQR as a more robust measure of spread. Matching at double the coarseness—i.e., one standard deviation—can eliminate more than 90% of the bias (Cochran 1968, Imbens 2004). We implement this far more stringent matching criterion to ensure that the failure curves follow each other as closely as possible.

[11] We control for these measures without estimating their trends because much of the service cost (in the warranty period) is driven by failures. Indeed, if we match on failures then our analysis (not presented here) indicates that, during the warranty period, the aggregate trends of the service cost measures across matched populations are not significantly different.

[12] The overall problem has 41,846 decision variables and 292,944 constraints; we solve it using the IBM ILOG CPLEX Optimization Studio.

[13] For the failure curve variables that are matched on a one-to-one level, we can alternatively analyze the distribution of pairwise differences. This procedure also shows that the means of the pairwise differences are small—i.e. less than 0.2 standard deviations away from zero.

[14] All of our results are robust to controlling for the matched variables (not reported here).

[15] In an exponential model, the translation from a coefficient δ to the effect size is given by $\exp\{\delta\} - 1$.

[16] We de-mean *CustomerSize* in the regression, so that the coefficient for δ can be interpreted as the incentive effect on an operator at the mean *CustomerSize*.

[17] We de-mean *Age* so that we can interpret the coefficient of δ as the incentive effect at the mean (as described for *CustomerSize* in Endnote 16).

[18] The observed increase in failure counts can arise due to three different reasons: first, operators on full-protection plans increase their usage levels; second, they tend to report more failures; and finally, they reduce the level of equipment care. Using data on equipment usage and failure severity (which we use as a proxy for misreporting), we show in the appendix that the first two factors do not vary significantly across contract types.

[19] Because we do not observe significant variation in failure severity across contract types (see Endnote 18), the higher on-site response rate is *not* driven by systematic variations in problem complexity but rather incentive effects.

References

Aarset MV (1987) How to identify a bathtub hazard rate. *IEEE Trans. Reliability* 36(1):106–108.

Abbring JH, Heckman JJ, Chiappori PA, Pinquet J (2003) Adverse selection and moral hazard In insurance: Can dynamic data help to distinguish? *J. Eur. Econom. Assoc.* 1(2–3):512–521.

Angrist JD, Pischke JS (2008) *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton University Press, Princeton, NJ).

Bajari P, Houghton S, Tadelis S (2014) Bidding for incomplete contracts : An empirical analysis of adaptation costs. *Amer. Econom. Rev.* 104(4):1288–1319.

Banerjee AV, Duflo E (2000) Reputation effects and the limits of contracting: A study of the Indian software industry. *Quart. J. Econom.* 115(3):989–1017.

Bhattacharya S, Lafontaine F (1995) Double-sided moral hazard and the nature of share contracts. *RAND J. Econom.* 26(4): 761–781.

Block HW, Savits TH (1997) Burn-in. *Statistical Sci.* 12(1):1–19.

Cachon GP (2003) Supply chain coordination with contracts. *Handbooks Oper. Res. Management Sci.* 11:227–339.

Cameron AC, Trivedi PK (2009) *Microeconometrics Using Stata* (Stata Press, College Station, TX).

Chiappori P-A, Salanié B (2003) Testing contract theory: A survey of some recent work. Dewatripont M, Hansen L, Turnovsky S, eds. *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress, Econometric Society Monographs* (Cambridge University Press, Cambridge, UK), 115–149.

Chisholm DC (1997) Profit-sharing versus fixed-payment contracts: Evidence from the motion pictures industry. *J. Law, Econom., Organ.* 13(1):169–201.

Cochran WG (1968) The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 24(2):295–313.

Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences* (Lawrence Erlbaum Associates, Hillsdale, NJ).

Corbett CJ, DeCroix GA, Ha AY (2005) Optimal shared-savings contracts in supply chains: Linear contracts and double moral hazard. *Eur. J. Oper. Res.* 163(3):653–667.

Corts KS, Singh J (2004) The effect of repeated interaction on contract choice: Evidence from offshore drilling. *J. Law, Econom., Organ.* 20(1):230–260.

Cui R, Allon G, Bassamboo A, Van Mieghem JA (2015) Information sharing in supply chains: An empirical and theoretical valuation. *Management Sci.* 61(11):2803–2824.

Deloitte Research (2006) *The Service Revolution in Global Manufacturing Industries* (Deloitte, New York).

ECRI Institute (2013) Healthcare product comparison system. Technical Report 610, ECRI Institute, Plymouth Meeting, PA.

Fuchs VR, Leveson IH (1968) *The Service Economy* (National Bureau of Economic Research, Cambridge, MA).

Gopal A, Sivaramakrishnan K, Krishnan MS (2003) Contracts in offshore software development: An empirical analysis. *Management Sci.* 49(12):1671–1683.

Guajardo JA, Cohen MA, Kim S-H, Netessine S (2012) Impact of performance-based contracting on product reliability: An empirical analysis. *Management Sci.* 58(5):961–979.

Helland E (2003) Contingency fees, settlement delay, and low-quality litigation: Empirical evidence from two data sets. *J. Law, Econom., Organ.* 19(2):517–542.

Iacus SM, King G, Porro G (2011) Causal inference without balance checking: Coarsened exact matching. *Political Anal.* 20(1):1–24.

Imai K, King G, Stuart EA (2008) Misunderstandings between experimentalists and observationalists about causal inference. *J. Roy. Statist. Soc. Ser. A Statist. Soc.* 171(2):481–502.

Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econom. Statistics* 86(1):4–29.

Jain N, Hasija S, Popescu DG (2013) Optimal contracts for outsourcing of repair and restoration services. *Oper. Res.* 61(6):1295–1311.

Kalnins A, Mayer KJ (2004) Relationships and hybrid contracts: An analysis of contract choice in information technology. *J. Law, Econom., Organ.* 20(1):207–229.

Karmarkar US, Roels G (2015) An analytical framework for value co-production in services. *Service Sci.* 7(3):163–180.

Kim S-H, Cohen MA, Netessine S (2007) Performance contracting in after-sales service supply chains. *Management Sci.* 53(12): 1843–1858.

Kreps DM (1997) Intrinsic motivation and extrinsic incentives. *Amer. Econom. Rev.* 87(2):359–364.

Levin J, Tadelis S (2010) Contracting for government services: Theory and evidence from U.S. cities. *J. Indust. Econom.* 58(3):507–541.

Nagarajan M, Sošić G (2008) Game-theoretic analysis of cooperation among supply chain agents: Review and extensions. *Eur. J. Oper. Res.* 187(3):719–745.

Ning J, Babich V, Handley J, Keppo J (2018) Risk-aversion and B2B contracting under asymmetric information: Evidence from managed print services. *Oper. Res.* 66(2):392–408.

Novak S, Stern S (2008) How does outsourcing affect performance dynamics? Evidence from the automobile industry. *Management Sci.* 54(12):1963–1979.

Novak S, Stern S (2009) Complementarity among vertical integration decisions: Evidence from automobile product development. *Management Sci.* 55(2):311–332.

Nowlan FS, Heap HF (1978) Reliability centered maintenance. Technical report, United Airlines, San Francisco.

Oliva R, Kallenberg R (2003) Managing the transition from products to services. *Internat. J. Service Indust. Management* 14(2): 160–172.

Plambeck EL, Zenios SA (2000) Performance-based incentives in a dynamic principal-agent model. *Manufacturing Service Oper. Management* 2(3):240–263.

Ramdas K, Randall T (2008) Does component sharing help or hurt reliability? An empirical study in the automotive industry. *Management Sci.* 54(5):922–938.

Roels G, Karmarkar US, Carr S (2010) Contracting for collaborative services. *Management Sci.* 56(5):849–863.

Rosenbaum PR, Ross RN, Silber JH (2007) Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *J. Amer. Statistical Assoc.* 102(477): 75–83.

Sawhney M, Balasubramanian S, Krishnan V (2004) Creating growth with services. *MIT Sloan Management Rev.* 45(2):1–17.

Smith SA, Oren SS (1980) Reliability growth of repairable systems. *Naval Res. Logist. Quart.* 27(4):539–547.

Smith-Bindman R, Miglioretti DL (2012) Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996–2010. *J. Amer. Medical Assoc.* 307(22):2400–2409.

Stuart EA (2010) Matching methods for causal inference: A review and a look forward. *Statistical Sci.* 25(1):1–21.

Susarla A (2012) Contractual flexibility, rent seeking, and renegotiation design: An empirical analysis of information technology outsourcing contracts. *Management Sci.* 58(7):1388–1407.

Susarla A, Subramanyam R, Karhade P (2010) Contractual provisions to mitigate holdup: Evidence from information technology outsourcing. *Inform. Systems Res.* 21(1):37–55.

Terwiesch C, Ren ZJ, Ho TH, Cohen MA (2005) An empirical analysis of forecast sharing in the semiconductor equipment supply chain. *Management Sci.* 51(2):208–220.

Wooldridge JM (2010) *Econometric Analysis of Cross Section and Panel Data* (MIT Press, Cambridge, MA).

Yang D, Small DS, Silber JH, Rosenbaum PR (2012) Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. *Biometrics* 68(2):628–636.

Zhou Y-P, Ren ZJ (2011) Service outsourcing. Cochran JJ, Cox LA Jr, Keskinocak P, Kharoufeh JP, Smith JC, eds. *Wiley Encyclopedia of Operations Research and Management Science* (John Wiley & Sons, New York).

Zubizarreta JR (2012) Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Amer. Statistical Assoc.* 107(500):1360–1371.

Zubizarreta JR, Paredes RD, Rosenbaum PR (2014) Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *Ann. Appl. Statistics* 8(1):204–231.