Optimizing the preventive maintenance frequency with causal machine learning

Toon Vanderschueren[a,b,*], Robert Boute[c,d,e], Tim Verdonck[b,f], Bart Baesens[a,g] and Wouter Verbeke[a]

[a] *Research Centre for Information Systems Engineering, Faculty of Economics and Business, KU Leuven, Leuven, Belgium*
toon.vanderschueren@kuleuven.be, bart.baesens@kuleuven.be, wouter.verbeke@kuleuven.be

[b] *Applied Mathematics, Department of Mathematics, University of Antwerp, Antwerp, Belgium*
tim.verdonck@uantwerpen.be

[c] *Research Centre for Operations Management, Faculty of Economics and Business, KU Leuven, Leuven, Belgium*
robert.boute@kuleuven.be

[d] *Technology and Operations Management Area, Vlerick Business School, Leuven, Belgium*

[e] *VCCM, Flanders Make, Belgium*

[f] *Statistics and Data Science, Department of Mathematics, KU Leuven, Leuven, Belgium*

[g] *Department of Decision Analytics and Risk, Southampton Business School, Southampton, United Kingdom*

ARTICLE INFO

Keywords:
Maintenance
Causal inference
Individual treatment effects
Machine learning

ABSTRACT

Maintenance is a challenging operational problem where the goal is to plan sufficient preventive maintenance (PM) to avoid asset overhauls and failures. Existing work typically relies on strong assumptions (1) to model the asset's overhaul and failure rate, assuming a stochastic process with known hazard rate, (2) to model the effect of PM on this hazard rate, assuming the effect is deterministic or governed by a known probability distribution, and (3) by not taking asset-specific characteristics into account, but assuming homogeneous hazard rates and PM effects. Instead of relying on these assumptions to model the problem, this work uses causal inference to *learn* the effect of the PM frequency on the overhaul and failure rate, conditional on the asset's characteristics, from observational data. Based on these learned outcomes, we can optimize each asset's PM frequency to minimize the combined cost of failures, overhauls, and preventive maintenance. We validate our approach on real-life data of more than 4,000 maintenance contracts from an industrial partner. Empirical results on semi-synthetic data show that our methodology based on causal machine learning results in individualized maintenance schedules that are more accurate and cost-effective than a non-causal approach that does not deal with selection bias and a non-individualized approach that prescribes the same PM frequency to all machines.

1. Introduction

Maintenance constitutes an intricate operational problem. The challenge is to avoid failures and costly overhauls, while simultaneously minimizing the cost of preventive maintenance (PM). We consider the problem of deciding on the frequency of PM interventions, where the optimal frequency minimizes the combined cost of both PM and detrimental outcomes resulting from deterioration, such as failures or overhauls. To optimize the PM frequency, existing work typically makes strong assumptions regarding the asset's *hazard rate*, i.e., the frequency with which failures and overhauls occur, and the *effect of PM* on this hazard rate. Moreover, existing maintenance policies assume asset homogeneity in the hazard rate and/or PM effect by not taking asset characteristics into account. We argue that all of these assumptions can be violated in practice.

First, most work assumes the asset's *overhaul and failure rates* follow a stochastic process that is known to the decision-maker, which is typically not the case in practice (de Jonge et al., 2015). Moreover, estimating the parameters of the stochastic process from data is challenging due to censoring (Louit et al., 2009; Fouladirad et al., 2018) and still requires assuming a certain type of statistical distribution that might not coincide with the actual overhaul or failure rate. Finally, most existing work assumes asset homogeneity and does not incorporate the effects of the asset's characteristics on the overhaul and failure rates. In reality, however, an older asset might be more prone to failure and require more PM interventions than a more recent one.

Existing work also requires assumptions on *the effect of PM* on the overhaul and failure rates. A broad spectrum of maintenance effects have been studied in the literature, ranging from perfect maintenance, which restores the system to a state as good as new, to worst maintenance, where maintenance causes the asset to fail (Pham and Wang, 1996). Existing approaches in imperfect maintenance assume that the effect is either deterministic or stochastic following a specified probability distribution. These assumed effects, however, might not always correspond to the actual effect. Moreover, the effect of PM is typically assumed to be identical for all assets. In reality, the effect of the same type of PM intervention could be very different for different assets. For example, changing a gear would likely have a different impact on a brand-new asset compared to the exact same maintenance intervention on an old, worn-down asset.

✉ toon.vanderschueren@kuleuven.be (T. Vanderschueren)
ORCID(s): 0000-0001-7202-1589 (T. Vanderschueren); 0000-0003-2890-2797 (R. Boute); 0000-0003-1105-2028 (T. Verdonck); 0000-0002-5831-5668 (B. Baesens); 0000-0002-8438-0535 (W. Verbeke)

In this work, we relax these assumptions regarding the hazard rate and the PM effect. Instead, we propose a data-driven maintenance policy that learns the effect of the PM frequency on the resulting overhaul and failure rates, conditional on the asset's characteristics. This approach allows to flexibly learn the outcomes for different PM frequencies from historical, observational data using machine learning, rather than assuming a prespecified (or known) hazard rate and PM effect based on expertise, and to design an asset-specific PM schedule based on the learned outcomes.

These benefits are achieved by framing maintenance as a problem of causal inference. We argue that the challenge in maintenance is that, for each specific asset, we only observe one overhaul and failure rate corresponding to the PM frequency that was administered in practice. We never observe the counterfactual outcomes, i.e., what would have happened if that asset had received more or less maintenance. Because of this, we never know whether the optimal PM frequency was prescribed. Causal inference offers a solution to this problem by predicting each individual asset's hypothetical overhauls and failures at different PM frequencies. By learning a model that predicts the overhaul and failure rate given the PM frequency, we can optimize the PM schedule to minimize the total estimated cost. Essentially, we propose using observational data to learn an asset-specific digital twin for maintenance that predicts the overhaul and failure rate should an asset be prescribed a certain PM frequency.

This work contributes to the extant literature on preventive maintenance by proposing a novel prescriptive framework for maintenance that prescribes each asset's desired preventive maintenance frequency based on the estimated effect of PM on its overhaul and failure rates. To this aim, we frame maintenance as a problem of causal inference and leverage state-of-the-art machine learning methods for causal inference. These models learn to estimate an asset's potential outcomes for different PM frequencies from observational data. Moreover, we formulate a prescriptive policy that uses the potential outcomes to decide on the optimal PM frequency that minimizes the total cost of failures, overhauls and PM interventions. Empirically, we contribute by demonstrating the use of our causal inference framework on a dataset consisting of more than 4,000 maintenance contracts of industrial equipment provided by an industrial partner. Finally, as our proposed approach itself comes with assumptions, we discuss their viability in the context of maintenance.

2. Related work

Maintenance has been studied extensively in operations research, with a wide variety of proposed maintenance policies (Wang, 2002; Ding and Kamaruddin, 2015; de Jonge and Scarf, 2020). Our work touches upon the literature on time-based maintenance, imperfect maintenance, condition-based maintenance, as well as prescriptive analytics and causal inference.

2.1. Time-based maintenance

We consider the problem of finding an optimal PM frequency, equivalent to finding the optimal period between PM interventions, known as time-based maintenance (Barlow and Hunter, 1960). This approach has been widely studied and, because of its simplicity, it is still frequently used in practice (Ahmad and Kamaruddin, 2012; Faccio et al., 2014). The key idea is to perform PM with a constant frequency throughout the asset's lifetime. Typically, this optimal PM frequency is found by modelling the stochastic overhaul and failure rates using a statistical distribution and then finding the PM frequency that minimizes the estimated total cost (Ahmad and Kamaruddin, 2012).

The drawback of most existing time-based maintenance policies is that they model failures and overhauls using an assumed stochastic process. Estimating the parameters of this stochastic process can be difficult due to censoring. This is because, in reality, assets are often maintained before failure occurs. Even if the parameters of the stochastic process can be estimated from data, the process itself can be misspecified. Moreover, existing work on time-based maintenance typically does not consider asset heterogeneity. Our proposed approach does not rely on a parametric model of the asset's overhauls and failures, but estimates each asset's overhaul and failure rates given the PM frequency using a flexible machine learning model, conditional on that asset's characteristics.

2.2. Imperfect maintenance

Most existing work assumes that preventive maintenance restores the system to a state that is as good as new. However, maintenance is typically imperfect in reality. Different maintenance effects have been studied in the literature, ranging from maintenance that restores the system to a perfect state to maintenance that makes the system's state worse (Pham and Wang, 1996). Consequently, developing maintenance policies that incorporate imperfect maintenance is an important research problem.

Existing work models the effect of imperfect maintenance as either stochastic (based on a known probability distribution) or deterministic (Pham and Wang, 1996; Chukova et al., 2004). Stochastic effects include the (p, q) rule, where maintenance is as good as new with probability p and as good as old with probability $q = 1 - p$ (Nakagawa, 1979a,b; Brown and Proschan, 1983), and its age-dependent variant $(p(t), q(t))$ (Block et al., 1985). Other work assumes a deterministic effect. Improvement factor models assume that maintenance decreases the system's failure rate by a deterministic improvement factor (Malik, 1979). Similarly, in virtual age models, imperfect maintenance decreases the system's age or failure rate with a deterministic factor q where $0 < q < 1$ (Kijima, 1989; Tanwar et al., 2014).

The literature has proposed methods for estimating the parameters of these imperfect maintenance models from data and corresponding goodness-of-fit tests (Liu et al., 2011; de Toledo et al., 2015; Zhang and Xie, 2017). However, these approaches still start from a (deterministic or stochastic) model of the PM effect that can be misspecified in practice. Moreover, the goodness-of-fit tests only verify whether the model corresponds to the asset pool globally. Conversely, our approach estimates an effect that is, first, model-free as it does not assume a certain type of effect and, second, machine-dependent, as it is based on individual characteristics. This is achieved by learning the overhaul and failure rates for different PM frequencies from observational data. Finally, a key difference with our approach is that we do not consider the effect of a single PM intervention, but rather focus on the outcomes over a period of time caused by different PM frequencies.

2.3. Condition-based maintenance

Data-driven, condition-based maintenance policies have recently gained importance in the maintenance literature (Bousdekis et al., 2021). Condition-based maintenance is a policy in which maintenance is optimized based on the machine's state or its characteristics (Gits, 1992; Alaswad and Xiang, 2017). Especially relevant to our work are recent, predictive maintenance approaches that learn a predictive model from data to decide on the appropriate maintenance interventions (Swanson, 2001; Carvalho et al., 2019). Various authors propose using neural networks due to their flexibility and ability to extract features from data (Fast et al., 2008; Tian, 2012; Wu et al., 2013; Lu et al., 2018).

A typical approach is to predict the machine's health from its characteristics and apply maintenance when a degradation threshold is reached. This is achieved by monitoring the machine's health using a data-driven model to predict whether a failure is imminent. When the perceived risk is too high, e.g., exceeding a degradation threshold, an intervention can be scheduled to avoid failure (e.g., as in Bey-Temsamani et al., 2009; Do et al., 2015; Matyas et al., 2017; Poppe et al., 2018; Nemeth et al., 2018; Ansari et al., 2019). Therefore, various works have proposed predicting failures using machine learning models (Kusiak and Verma, 2011; Lee et al., 2017; Leukel et al., 2021) with several recent approaches based on neural networks specifically (Jansen et al., 2018; Chen et al., 2019a,b, 2020; Savitha et al., 2020; Orrù et al., 2020; Alves et al., 2020; Zhao and Huang, 2021; Ye and Yu, 2021; Figueroa Barraza et al., 2022). By incorporating asset characteristics in the estimation, predictive policies can account for asset heterogeneity.

The downside of condition-based approaches is that they only predict the asset's deterioration and do not consider the impact of PM on this deterioration. The time at which the deterioration threshold is reached and PM is planned, might not correspond to the optimal timing to most effectively perform maintenance and remedy the deterioration. Ideally, maintenance should not be performed just before the asset fails, but at the time when it is most effective at lowering the asset's failure probability. To this end, it is important to estimate the asset's hazard rate *resulting from* a certain PM frequency, which is exactly what our approach aims to achieve.

Similar to the general literature on imperfect maintenance, existing condition-based approaches that consider imperfect maintenance also assume either a deterministic or stochastic maintenance effect. There exist three broad categories of condition-based approaches that account for imperfect maintenance (Alaswad and Xiang, 2017). A first category considers minimal maintenance with a *deterministic* effect, in which a system has several deterioration stages and imperfect maintenance returns the system to the previous stage. A second category considers *stochastic* effects, where the maintenance effect is governed by an assumed probability distribution. Finally, in *improvement factor* models, imperfect maintenance decreases the system's hazard rate with a (deterministic) factor between zero and one. To the best of our knowledge, no existing condition-based approaches aim to *learn* the effect of maintenance from data.

2.4. Prescriptive analytics and causal inference

Instead of assuming a hazard rate or PM effect, this work uses machine learning models to learn the effect of maintenance using techniques from causal inference. Causal inference aims to estimate the effect of a certain cause from data, in our case the failure rate resulting from a given PM frequency. Ideally, estimating maintenance effects would be done by conducting a randomized controlled trial: assigning different PM frequencies to a collection of (similar) machines and comparing the outcomes (Rubin, 1974). However, in practice, this approach can be prohibitively expensive, unfeasible, or even unethical (e.g., when considering life support equipment in hospitals). In maintenance, it would generally be challenging to randomly assign varying levels of PM to different machines, because an excessively low PM frequency might risk not ensuring minimal service levels. When randomized controlled trails are impossible, we need to rely on historical, observational data of machines and their maintenance to learn the outcomes caused by different PM frequencies.

The challenge of working with observational data is that this data is biased due to existing maintenance policies that were in use (Rubin, 1974). For example, as a result of an existing policy, machines more prone to failure might have been more likely to receive more maintenance in the past. This phenomenon, called selection bias or confounding bias, can result in biased estimates of the counterfactual outcomes if ignored. Under certain assumptions, specialized tools from the causal inference literature can be used to tackle exactly this problem and learn causal effects from observational data, i.e., in the presence of selection bias (Yao et al., 2021). Specifically, our work is related to learning potential outcomes for continuous-valued interventions (Imbens, 2000; Hirano and Imbens, 2004; Imai and Van Dyk, 2004; Schwab et al., 2020; Bica et al., 2020), e.g., the PM frequency[1]. Learning the outcomes for different levels of a continuous treatment is also referred to as learning a dose-response curve.

[1] The number of PM interventions is discrete, but the number of PM interventions per running period (i.e., PM frequency) is continuous-valued.

Causal inference has been applied to a variety of applications, such as personalized medicine (Berrevoets et al., 2020), economic policy design (Athey and Wager, 2021), marketing (Varian, 2016; Devriendt et al., 2018), and education (Webbink, 2005). Moreover, it is related to prescriptive analytics (Verbeke et al., 2020, 2022), which has recently gained importance in operations research (Bertsimas et al., 2019; Bertsimas and Kallus, 2020). This work uses causal inference to predict a machine's failure rate and overhaul rate for different PM frequencies and, consequently, to decide upon a personalized PM schedule. To the best of our knowledge, this is the first application of causal inference for maintenance optimization.

3. Problem overview

This work aims to solve the problem of prescribing an asset's PM frequency to minimize the costs resulting from overhauls, failures, and PM. In particular, we are motivated by the challenge faced by a provider of full-service maintenance contracts. The service provider is responsible for maintaining the client's asset at a predetermined price (Deprez et al., 2021). To maximize its profit margin, the service provider needs to decide on the PM frequency that minimizes the costs of failures, overhauls, and PM. The PM frequency is usage-based and defined over a running period, which corresponds to a standardized number of running hours. For each contract, the service provider has access to contract characteristics, such as the type of machine it concerns and the machine's age at contract start. We consider each machine as a single-unit system.

We assume the service provider conducts a single type of planned PM intervention and needs to decide on the frequency of these interventions. Planned PM aims to prevent two types of events: overhauls and failures. The first, *overhauls*, are unplanned, comprehensive maintenance interventions during which large parts of the machinery need to be replaced. From the viewpoint of the full-service maintenance provider, these are the most costly type of event. The second, machine *failures*, are also unplanned and result in an urgent need for maintenance as the machine stops running until corrective maintenance occurs. A failure also incurs a cost to the service provider that is smaller than the cost of an overhaul, but larger than the cost of PM.

The overall goal is to find each contract's optimal PM frequency that minimizes the combined cost of planned PM, overhauls and failures, from the perspective of the service provider. Although planning more PM interventions is likely to result in less overhauls and failures, it comes at an increased maintenance cost. Therefore, the optimal PM frequency is a trade-off between costs resulting from planned PM on the one hand and costs resulting from unplanned overhauls and failures on the other hand. Due to heterogeneity in the contracts and associated machines, maintenance might need to be planned more frequently for some contracts. Therefore, it is important to consider the contract's characteristics when deciding on the PM frequency. To this aim, the service provider has access to information on past contracts, including the administered PM frequency, and the overhaul and failure rates observed for that PM frequency.

Let each contract be defined as a tuple $\big(\mathbf{X}, T, O(T), F(T)\big)$. $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$ denotes a vector of (static) characteristics of the contract and the associated machine. The treatment, in our case the PM frequency, corresponds to the number of PM interventions that were applied per running period, and is denoted as $T \in \mathcal{T} \subset \mathbb{R}^+$. Finally, $O \in \mathcal{O} \subset \mathbb{R}^+$ and $F \in \mathcal{F} \subset \mathbb{R}^+$ are the observed number of overhauls and failures per running period, i.e., the overhaul and failure rates. Following the causal inference literature, we adopt the Rubin–Neyman potential outcomes framework (Rubin, 2004, 2005) and denote the overhaul intensity O and failure rate F for a given maintenance frequency t as $O(t)$ and $F(t)$.

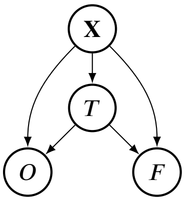


Figure 1: Diagram illustrating the assumed causal relationships between the different variables. X: Asset characteristics, T: PM frequency, O: Overhaul rate, and F: Failure rate.

The objective is to decide on each asset's optimal PM frequency t_i^* that minimizes the total cost per running period. We assume a cost model per running period similar to Faccio et al. (2014). Each asset i's cost per running period consists of the combined costs of PM, overhauls and failures, which all depend on the decision-variable, i.e., the PM frequency t_i:

$$c_i(t_i) = \underbrace{c_t\, t_i}_{\text{PM}} + \underbrace{c_o\, o_i(t_i)}_{\text{Overhauls}} + \underbrace{c_f\, f_i(t_i)}_{\text{Failures}}, \tag{1}$$

for $i \in \{1, \ldots, n\}$. We assume that the costs of PM, overhauls, and failures are deterministic and known ($c_t, c_o, c_f \in \mathbb{R}^+$).

To assist the service-provider's decision-making, data is available on m past contracts $\mathcal{D} = \big\{(\mathbf{x}_i, t_i, o_i, f_i)\big\}_{i=1}^n$. For each of these past contracts, only one potential outcome was observed for O and F given that contract's PM frequency T: $o_i(t_i)$ and $f_i(t_i)$. The other, counterfactual outcomes are never observed—this is known as the fundamental problem of causal inference (Holland, 1986). The challenge in causal inference is to predict, for a new contract, the potential outcomes for all possible values of T using historical, observational data.

For each observed contract i, decisions regarding the administered PM frequency t_i were based on its characteristics $\mathbf{x}_i$ according to a (possibly unknown) existing policy, resulting in selection bias or confounding bias in the data. In observational data, we can expect a relationship between an asset's characteristics and the PM frequency it received. For example, the service provider might know from experience that older machines are more likely to fail when not receiving frequent PM and, because of this, typically prescribed more maintenance to those machines in the past. Factors that influence both the administered PM frequency and the outcome, the failure and overhaul rate, are called confounders. In this example, age is a confounder affecting both the received PM frequency and the resulting failure rate. We show the assumed causal structure of the problem in Figure 1.

The presence of confounders and selection bias is typically the case when working with observational data. This is because past PM frequencies were not assigned at random, but based on information on the contract and machine. Because of the associations between confounders and the PM frequency, assets that received relatively infrequent PM are different from assets that received relatively more frequent PM. This phenomenon, called selection bias, complicates learning the relationships between overhaul and failure rates, the PM frequency, and asset characteristics. Therefore, when learning a predictive model for estimating the overhaul and failure rate resulting from a given PM frequency from observational data, we are required to adjust for selection bias to obtain unbiased estimates.

4. Methodology

Our methodology consists of a predict-then-optimize framework, see Figure 2 for a high-level overview. First, we predict each new contract's potential outcomes, i.e., its overhaul $o_i(t)$ and failure rate $f_i(t)$ for PM frequencies $t \in T$, based on its characteristics $\mathbf{x}_i$. Therefore, the first step is to train a machine learning model to estimate these potential outcomes from observational data on past contracts $\mathcal{D}$. In a second phase, we use these predictions to estimate each contract's total cost per running period for different PM frequencies $t \in T$. The PM frequency is chosen to minimize the resulting total cost of overhauls, failures, and PM.

In what follows, we first introduce standard assumptions that are required to estimate potential outcomes from observational data in Section 4.1. Second, we describe how we estimate the potential outcomes by learning a causal machine learning model from observational data. We used a state-of-the-art methodology called SCIGAN (Bica et al., 2020). This is described in Section 4.2. Third, in Section 4.3, we describe how these predictions are used to determine each machine's optimal PM frequency that minimizes the total estimated cost.

4.1. Assumptions

The challenge in estimating potential outcomes from observational data is dealing with selection bias. Learning unbiased estimates of the potential outcomes from observational data requires making three standard assumptions: consistency, overlap, and unconfoundedness (Imbens, 2000; Bica et al., 2020). The first, *consistency*, means that each contract's observed outcomes for O and F given PM frequency $T = t$ are its potential outcomes $O(t)$ and $F(t)$:

Assumption 1: Consistency. *$Y = Y(t)$ for all $t \in T$.*

This assumption implies that there is only one version of the treatment and that the mechanism used to assign the treatment does not matter. It is violated if, for example, the prescribed PM is not performed for some assets, e.g., when some clients do not adhere to the PM frequency prescribed by the service provider. Consistency may seem straightforward, but ensures that the PM schedule prescribed by the service provider will be observed in practice.

The second assumption, *overlap or positivity*, ensures that each possible contract $\mathbf{x}_i$ has a non-zero probability of receiving each frequency of PM interventions t_i:

Assumption 2: Overlap. *For all $\mathbf{x} \in \mathcal{X}$ with $p(\mathbf{x} > 0)$ and $t \in \mathcal{T} : 0 < p(t|\mathbf{x}) < 1$.*

This implies that, for each observed machine, it was a priori possible to observe each PM frequency, although not necessarily with the same probability. This assumption would be violated when, for example, machines older than five years always receive at least ten PM interventions per running period. In that case, the probability of receiving a PM frequency lower than ten is zero for those machines, implying a violation of the overlap assumption. In that case, we would not be able to account for selection bias, as no observations would exist to infer what would happen to old machines at low PM frequencies.

The third and final assumption, *unconfoundedness or no hidden confounders*, ensures that there are no unobserved variables influencing both the treatment assignment T and a potential outcome $O(t)$ or $F(t)$:

Assumption 3: Unconfoundedness. *Conditional on machine characteristics $\mathbf{X}$, potential outcomes $O(t)$ and $F(t)$ are independent of the PM frequency T:*

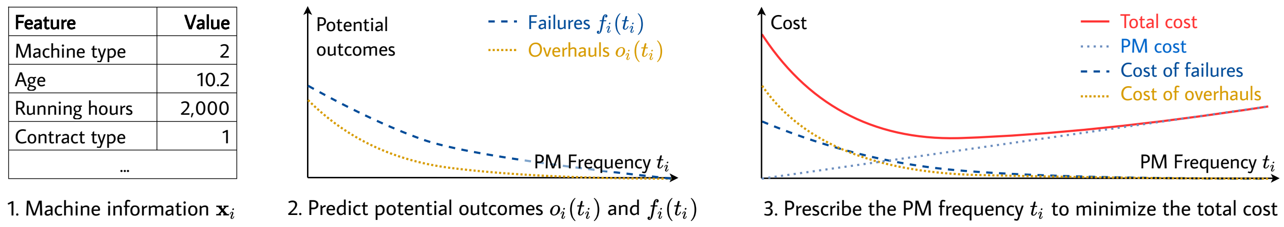
$$\{O(t), F(t) | t \in \mathcal{T}\} \perp\!\!\!\perp T | \mathbf{X}.$$

This assumption implies that all information that informed decisions regarding past PM frequencies are included in the data. This assumption would be violated if, for example, machines in some locations were maintained more frequently in the past, but no record of the machine's location was kept. If hidden confounders are present, it is impossible to adjust for the hidden confounder and, consequently, for selection bias based on the observed data. Given that Assumptions 1 to 3 are met, controlling for confounding resulting from machine characteristics $\mathbf{x}_i$ allows accounting for selection bias in observational data and obtain unbiased estimates.

4.2. Predictive model to estimate the effect of the PM frequency on overhaul and failure rates

First, we need to predict each contract's potential outcomes, i.e., its overhaul $o_i(t)$ and failure rate $f_i(t)$ for each PM frequency $t \in T$, based on its characteristics $\mathbf{x}_i$. To this aim, we learn two machine learning models $g_o : \mathcal{X} \times \mathcal{T} \rightarrow \mathcal{O}$ and $g_f : \mathcal{X} \times \mathcal{T} \rightarrow \mathcal{F}$ defined by parameters $\theta_o, \theta_g \in \Theta$. The goal is to obtain unbiased estimators of the potential outcomes:

$$g_o(t, \mathbf{x}) = \mathbb{E}\left[O(t)|\mathbf{X} = \mathbf{x}\right], \tag{2}$$



1. Machine information $\mathbf{x}_i$ 2. Predict potential outcomes $o_i(t_i)$ and $f_i(t_i)$ 3. Prescribe the PM frequency t_i to minimize the total cost

Figure 2: Methodology overview. We present a high-level overview of our methodology. First, machine characteristics $\mathbf{x}_i$ are used to predict the potential outcomes in terms of overhauls $o_i(t)$ and failures $f_i(t)$. Based on these estimates, the total cost for different PM frequencies $t \in T$ can then be estimated. Finally, the PM frequency $\hat{t}_i^*$ is chosen to minimize the total expected cost.

$$g_f(t, \mathbf{x}) = \mathbb{E}\left[F(t)|\mathbf{X} = \mathbf{x}\right]. \tag{3}$$

In this work, we learn g_o and g_f using SCIGAN, a recently proposed methodology for predicting potential outcomes of continuously-valued treatments that achieved state-of-the-art performance across a variety of settings (Bica et al., 2020). Each model is learned in two steps. First, a generative adversarial network (GAN) (Goodfellow et al., 2020) is trained to model the distribution of the potential outcomes, conditional on the contract's characteristics. This is achieved by training two neural networks, where the generator network learns to generate counterfactual contracts that cannot be distinguished from factual, observed contracts by the discriminator network. In a second phase, the GAN is used to augment the observed training data with generated counterfactual samples. This way, the augmented data set contains all potential outcomes, including the factual outcome and the generated, counterfactual outcomes. This way, the fundamental problem of causal inference is alleviated as we "observed" all potential outcomes for each contract and, because of this, the augmented data set does not suffer from selection bias. Using the augmented data set, a predictive model can be trained to predict the potential outcomes in a supervised manner. For this, we again use a neural network. Each network is implemented as a multilayer perceptron (MLP). Section A provides more information on the training and hyperparameter optimization of the models.

4.3. Optimization of the maintenance cost

The predicted potential outcomes allow estimating the costs incurred at different PM frequencies. It can be seen that, using the predicted potential outcomes, all terms in Equation (1) depend on the PM frequency t_i:

$$c_i(t_i) = c_t\, t_i + c_o\, o_i(t_i) + c_f\, f_i(t_i). \tag{4}$$

Each machine's optimal PM frequency t_i^* is found by minimizing the expected cost: $t_i^* = \text{argmin}_{t_i}\, c_i(t_i)$ for all $i \in \{1, \ldots, n\}$. To account for heterogeneity in the contracts, the PM frequency is optimized for each specific machine.

5. Results

We validate our methodology empirically using data provided by an original equipment manufacturer that offers full-service maintenance contracts to their customer base. By optimizing the PM frequency, they can minimize the total cost of such a contract, resulting from PM, overhauls, and failures. In Section 5.1, we first present the data used in our experimental analysis. Section 5.2 describes the semi-synthetic data generating procedure that we used to evaluate the predicted potential outcomes and prescribed PM frequencies. In Section 5.3, we present the evaluation metrics and benchmarks used. Finally, Section 5.4 presents the empirical results of our experimental analysis.

5.1. Data

Our data set contains more than 4,000 full-service maintenance contracts. For each contract i, we have information $\mathbf{x}_i$ relating to the characteristics of the machine, the contract, and maintenance-related events. An overview of the information available in the data is presented in Table 1 and an excerpt is shown in Table 2. Maintenance-related events (PM interventions, overhauls, and failures) are presented per running period, which is a set number of running hours. For reasons of confidentiality, the exact number of running hours per period is not revealed in this article. Costs are averaged over all events and re-scaled for reasons of confidentiality.

The data is preprocessed as follows. Categorical variables are encoded with dummies and $\mathbf{x}_i$ is standardized. The PM interventions, overhauls, and failures that occurred throughout the contract are converted to the number of events per running period to calculate each contract's PM frequency, overhaul rate, and failure rate. For future contracts, the exact number of running hours might not be known when the contract starts, but an estimate would typically be available.

5.2. Semi-synthetic data generating procedure

In order to obtain a good predictive model, we need to be able to accurately predict the overhaul and failure rates at different PM frequencies. However, if we test this predictor's accuracy using only observational data, we can verify the model's ability to accurately predict the observed outcome, the overhaul and failure rates only at the observed PM frequency t_i (the observed outcome), but not the overhaul and failure rates if the machine had received more or less maintenance (the unobserved outcomes). This makes the evaluation of causal models challenging, as only one potential outcome is observed for each contract in our dataset. Therefore, we rely on semi-synthetic data to evaluate our model. This

Variable	Domain
Machine information	
Type	$\{1, \dots, 7\}$
Age at contract start (in years)	$[0, 39]$
Running hours at contract start	$[2500, 110000]$
Contract information	
Type	$\{1, 2\}$
Duration (in days)	$[180, 5850]$
Running hours during contract	$[0, 186000]$
Average running hours per year	$[300, 8500]$
Preventive maintenance per running period	
PM frequency	$[0, 20]$
Outcomes per running period	
Number of overhauls	$[0, 128]$
Number of failures	$[0, 185]$
Average costs (in €)	
Preventive maintenance	73
Overhaul	207
Failure	104

Table 1
Data overview. Overview of the available contract information on machine and contract characteristics, preventive maintenance interventions, overhauls, and failures. For confidentiality, we present PM interventions, overhauls, and failures per running period, which is an undisclosed number of running hours. Similarly, the costs are averaged and re-scaled.

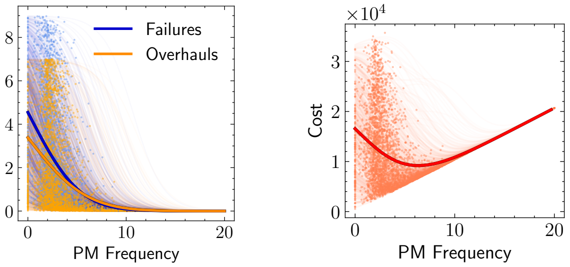
approach is commonly used in both maintenance (see e.g., Deprez et al., 2021) and causal inference (e.g., Berrevoets et al., 2020).

The key idea of the semi-synthetic setup is to create a test set containing each contract's potential outcomes at all possible PM frequencies, instead of only the observed outcome at one administered PM frequency. This is achieved by generating the outcomes at all possible PM frequencies $o_i(t)$ and $f_i(t)$ for all possible PM frequencies $t \in T$, based on the contract's real characteristics $\mathbf{x}_i$. This allows us to create (1) a training set with only one observed PM frequency for each contract, equivalent to observational data, and (2) a test set containing potential outcomes for all possible PM frequencies for each contract, which are never observed in reality but needed for evaluation.

Potential outcomes $o_i(t)$ and $f_i(t)$ are generated based on the observed characteristics $\mathbf{x}_i$ and PM frequencies $t \in T$. For the failure rates, we have:

$$f_i(t) = 9\,\sigma\Bigg(\underbrace{\mathbf{v}_f^\intercal \mathbf{x}_i}_{\text{Base rate}} - \underbrace{\frac{1}{10}\,\sigma\left(\mathbf{w}_f^\intercal \mathbf{x}_i\right) t}_{\text{PM effect}} + \underbrace{\epsilon_f}_{\text{Noise}} \Bigg), \quad (5)$$

with $\mathbf{v}_f, \mathbf{w}_f \sim \mathcal{U}\left((0,1)^{d\times 1}\right)$ and $\epsilon_f \sim \mathcal{N}(0,1)$. σ denotes the logistic function. This way, each machine has a base failure rate that is diminished by administering more frequent PM, where both the base rate and PM effect depend on the contract's characteristics $\mathbf{x}_i$. The factor 9 rescales the average failure rate to roughly the same number in the original, observed data. For the overhaul rates, we similarly have:



(a) Potential outcomes (b) Total cost

Figure 3: Semi-synthetic data. We represent the observed outcomes for contracts in the training and validation set by dots and the potential outcomes for contracts in the test set by a line. The bold lines illustrate the overhaul rate, failure rate, and total cost averaged across all contracts.

$$o_i(t) = 7\,\sigma\Bigg(\underbrace{\mathbf{v}_o^\intercal \mathbf{x}_i}_{\text{Base rate}} - \underbrace{\frac{1}{10}\,\sigma\left(\mathbf{w}_o^\intercal \mathbf{x}_i\right) t}_{\text{PM effect}} + \underbrace{\epsilon_o}_{\text{Noise}} \Bigg), \quad (6)$$

where $\mathbf{v}_o, \mathbf{w}_o \sim \mathcal{U}\left((0,1)^{d\times 1}\right)$ and $\epsilon_o \sim \mathcal{N}(0,1)$.

Using the semi-synthetic setup, the contracts in the test set have known potential outcomes for all possible values of $t_i \in \mathcal{T}$ based on Equations (5) and (6). Conversely, the training and validation sets include only one observed outcome for one PM frequency t_i. An illustration of a generated data set is shown in Figure 3. The training, validation, and test sets respectively consist of 50%, 25% and 25% of the data. Experiments are repeated five times.

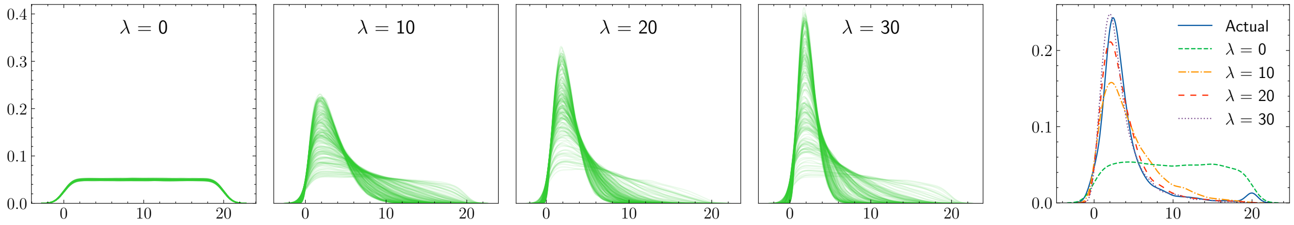
In a first analysis, we use the PM frequency t_i that was observed in practice for the training and validation set. In other words, we only simulate the overhaul and failure rates. In a subsequent analysis, we evaluate our approach for different levels of selection bias by also controlling the observed PM frequencies in the training and validation set. For this, we manipulate the level of selection bias by making the observed PM frequencies t_i more or less dependent on the contract characteristics $\mathbf{x}_i$, using an approach similar to Bica et al. (2020). More specifically, we control the selection bias by assigning PM frequencies based on sampling from a beta distribution:

$$t_i \sim 20\,\text{Beta}\left(1 + \frac{\lambda\delta_i}{10}, 1 + \lambda\delta_i\right), \quad (7)$$

where $\delta_i = \sigma(\mathbf{w}_b \mathbf{x}_i)$ with $\mathbf{w}_b \sim \mathcal{U}\left((0,1)^{d\times 1}\right)$. δ_i ensures that assignment of the PM frequency is based on observed features $\mathbf{x}_i$. This way, we control the level of selection bias by setting λ. A value of $\lambda = 0$ results in Beta(1, 1) or a uniform distribution. This implies that we randomly assign each machine's PM frequency with equal probability for each PM frequency in T. Therefore, $\lambda = 0$ results in a situation equivalent to a randomized controlled trial. Higher values of λ imply more selection bias, with $\lambda = 30$ resulting

Machine information			Contract information				Maintenance-related event frequencies		
Type	Age [years]	Running hours *contract start*	Type	Duration [days]	Running hours *during contract*	Running hours *average per year*	PM frequency t_i	Overhaul rate $o_i(t_i)$	Failure rate $f_i(t_i)$
4	0	528.88	1	1,826	12,391.63	2,434.67	1.42	0.19	0.91
5	12	77,301.37	1	1,764	29,131.68	4,907.42	2.24	2.57	7.24
6	15	39,312.72	1	2,555	8,906.65	2,694.49	2.89	5.92	8.47
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	16	61,948.75	0	1,764	21,303.56	3,912.07	4.02	0.25	2.52

Table 2
Data excerpt. We present an excerpt of the data set, showing examples of covariates related to the machine and contract $\mathbf{x}_i$, and maintenance related events per running period: the observed PM frequency t_i, the overhaul rate $o_i(t_i)$, and the failure rate $f_i(t_i)$.



(a) Effect of λ on the individual machines' PM frequency distributions

(b) PM frequencies

Figure 4: Simulating selection bias. (4a) We simulate a training data set where each contract's (observed) PM frequency is drawn from its probability distribution, shown in green, using Equation (7) for different values of λ. When $\lambda = 0$, each contract has equal probabilities of receiving each PM frequency between 0 and 20, corresponding to randomly assigned PM frequencies. Increasing λ makes the distributions more dependent on contract characteristics and therefore more diverse. This way, certain contracts will more likely receive less frequent PM, resulting in selection bias. Higher values of λ imply more diversity in the distributions and, consequently, more selection bias. (4b) We show how the PM frequency is distributed among the different contracts, both in reality and as a result of different values of λ. Larger values of λ result in more selection bias, with a value of 30 resulting in an overall PM frequency distribution close to the original.

in an overall distribution of the PM frequencies over the entire training set that is similar to the observed distribution. In other words, $\lambda = 30$ corresponds to a realistic level of selection bias. Figure 4a shows each contract's distribution from which the PM frequency is sampled, for different values of λ. A higher value of λ increases the diversity of the different contracts' PM frequency distributions, resulting in more selection bias in the training data. Figure 4b compares the observed distribution of PM frequencies over all contracts in the original data and the overall distributions of PM frequencies resulting from different values of λ.

5.3. Performance evaluation

We evaluate our predict-then-optimize approach using three different metrics. First, we evaluate the ability of the machine learning model to accurately predict a contract's overhaul $o_i(t)$ and failure rate $f_i(t)$ over different levels of PM frequencies $t \in T$. This is measured using the mean integrated square error (MISE) (Silva, 2016; Schwab et al., 2020):

$$\text{MISE} = \frac{1}{n}\sum_{i=1}^{n}\int_{0}^{m}\left(y_i(t) - \hat{y}_i(t)\right)^2 \,\mathrm{d}t, \tag{8}$$

for $y_i(t) \in \{o_i((t), f_i(t)\}$. Because we simulate the outcomes (see Figure 3), we know the ground truth $y_i(t)$ for each $t \in T$. Second, to evaluate the accuracy of the prescribed maintenance frequencies $\hat{t}_i^*$, we consider a variant of the policy error (PE) (Schwab et al., 2020) that compares the prescribed PM frequency with the optimal PM frequency:

$$\text{PE} = \frac{1}{n}\sum_{i=1}^{n}\left(t_i^* - \hat{t}_i^*\right)^2. \tag{9}$$

The optimal PM frequency t_i^* can be found numerically by searching over the total cost incurred at each possible PM frequency $t \in T$. Third, we evaluate the prescribed maintenance frequency in terms of costs using the policy cost ratio (PCR) that compares the costs of the estimated optimal PM frequency $c_i(\hat{t}_i^*)$ with the cost of the optimal PM frequency $c_i(t_i^*)$:

$$\text{PCR} = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i(\hat{t}_i^*)}{c_i(t_i^*)}. \tag{10}$$

For all metrics, a lower value indicates better performance with 0 being the optimal value for MISE and PE and 1 for PCR.

Our proposed maintenance policy uses SCIGAN to learn the individual treatment effects (ITE), i.e., each contract's overhaul and failure rate for different PM frequencies and will be referred to as SCIGAN–ITE. We benchmark against two other policies (see Table 3). First, a policy based on a neural network (MLP) that learns o_i and f_i given $\mathbf{x}_i$ and

Methodology	Selection bias?	Individualized?
SCIGAN–ITE	✓	✓
MLP–ITE	✗	✓
SCIGAN–ATE	✓	✗

Table 3
Methodologies overview. Our proposed, individual policy, SCIGAN–ITE, prescribes the PM frequency based on the individual treatment effect (ITE) estimated using SCIGAN. This proposed approach is analyzed using an ablation study and compared with two variants. The first, MLP–ITE, does not account for selection bias. The second, SCIGAN–ATE, is a general policy based on the average treatment effect (ATE) and is not individualized towards each individual machine.

t_i in a completely supervised manner without adjusting for selection bias (MLP–ITE). This allows us to assess whether there is a benefit of using the GAN to adjust for selection bias. Second, the average policy (SCIGAN–ATE) sets a single optimal $\hat{t}^*$ for all contracts based on the average (instead of the individual) PM effect. This allows to validate the benefit of an individualized policy tailored towards each specific machine.

5.4. Empirical results

In this section, we present the results of the semi-synthetic experiments based on more than 4,000 maintenance contracts. Section 5.4.1 addresses (1) whether there is improved performance resulting from adjusting for selection bias and (2) whether an individualized policy per contract outperforms a general policy that does not take contract characteristics into account. In Section 5.4.2, we show the importance of accounting for selection bias by evaluating the different policies' performance for varying levels of selection bias (Section 5.4.2).

5.4.1. Benefits of a causal, individualized PM policy

Table 4 reports the empirical results for the predictions and PM frequencies obtained using each methodology. The left part of Table 4 compares the ability of SCIGAN and MLP of accurately predicting the overhaul and failure rate at different PM frequencies. SCIGAN achieves the lowest error measured by the MISE. It predicts the overhaul and failure rate more accurately than the supervised MLP that does not account for selection bias. In the right part of Table 4, we assess the quality of the PM frequencies prescribed by the different approaches. These results show that the relatively more accurate predictions of the individualized, prescriptive approach (SCIGAN–ITE) also result in better PM frequencies. On the one hand, SCIGAN–ITE prescribes PM frequencies that are closer to the optimal PM frequency compared to the supervised (MLP–ITE) and non-individualized approach (SCIGAN–ATE), measured using the PE. On the other hand, SCIGAN–ITE also results in the lowest total cost as indicated by the PCR, achieving a cost that is 7% higher than the optimal policy, compared to 11% for MLP–ITE and 24% SCIGAN–ATE. The gap of 7% between SCIGAN–ITE and the optimal policy can be explained by the model being trained on limited data and the presence of noise in the data.

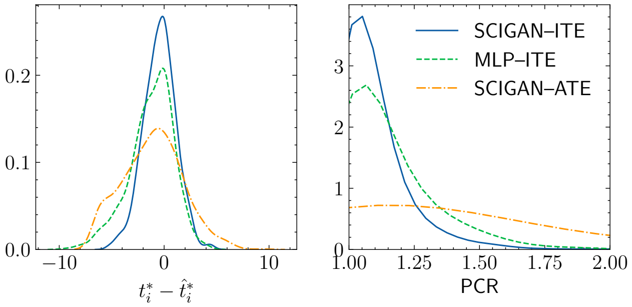


Figure 5: Evaluating the policies' decisions. We compare the accuracies and costs of the prescribed PM frequencies by looking at each model's performance over all contracts. (Left) We show how the differences between the prescribed and optimal PM frequency are distributed per model. (Right) We show the distribution of all contracts' policy cost ratios resulting from each model. Results are shown for one representative iteration.

Figure 5 takes a closer look at these results, by showing how each model's performance of each contract individually, rather than looking only at the average performance over all contracts. The left panel in Figure 5 assesses how close each contract's PM frequency is to the optimal PM frequency, by showing each model's error distribution, i.e., the differences between the prescribed and optimal PM frequencies for all contracts. For SCIGAN–ITE most of the errors are close to zero, indicating that the prescribed PM frequency is typically reasonably close to the optimal PM frequency. By comparison, MLP–ITE and SCIGAN–ATE more frequently prescribe a PM frequency that deviate from the optimal PM frequency, illustrated by the heavier tails in their distributions. The right panel in Figure 5 looks at the costs resulting from each contract's prescribed PM frequencies, by showing the distribution of each contract's PCR, i.e., the cost resulting from the prescribed PM frequency relative to the cost incurred by the optimal PM frequency. SCIGAN–ITE typically frequently obtains a PCR close to one, indicating that it incurs costs that are close to the optimal policy. By comparison, MLP–ITE and especially SCIGAN–ITE more frequently incur costs that are much higher than the costs resulting from the optimal PM frequency. These findings correspond to the findings averaged over all contracts in Table 4.

The improved performance of SCIGAN compared to a standard MLP suggests that learning PM effects from observational data requires accounting for selection bias. Moreover, the relatively worse performance of the non-individualized approach, SCIGAN–ATE, compared to the individualized approach, SCIGAN–ITE, shows the benefit of an individualized, machine-dependent policy for imperfect maintenance that takes into account machine characteristics and accounts for machine heterogeneity.

	MISE	
	Overhauls	Failures
SCIGAN	**7.71 ± 0.60**	**14.16 ± 1.68**
MLP	10.25 ± 1.33	18.27 ± 3.65

	PE	PCR
SCIGAN–ITE	**2.40 ± 0.46**	**1.07 ± 0.01**
MLP–ITE	4.36 ± 1.25	1.11 ± 0.02
SCIGAN–ATE	8.77 ± 1.07	1.24 ± 0.04

Table 4
Empirical evaluation. We compare performance for the different policies over five simulation runs. We evaluate each model's ability to accurately predict the potential outcomes $o_i(t)$ and $f_i(t)$ using the MISE, as well as each model's ability to accurately prescribe PM frequencies (PE) and to minimize costs (PCR). For all metrics, a lower value is better.

5.4.2. Importance of accounting for selection bias

The results in the previous section were obtained for the level of selection bias that was observed in reality, by using the PM frequencies in the training set as originally observed. In this section, we obtain more insight into the influence of selection bias by comparing the performance of SCIGAN–ITE and MLP–ITE for varying levels of selection bias. This is achieved by controlling the level of selection bias using λ (see Equation (7)). At $\lambda = 0$, there is no selection bias. In this case, each contract's PM frequency is randomly drawn from the domain of all possible PM frequencies, with each contract having equal probabilities of receiving each PM frequency. In other words, setting $\lambda = 0$ results in data similar to a randomized controlled trial, which would be ideal for learning causal effects. Even though randomly assigning PM frequencies is not reasonable in the context of maintenance, this simulation allows us to study the influence of selection bias on performance. Increasing λ makes a machine's observed PM frequency less random and more dependent on its characteristics and, therefore, results in more selection bias.

Figure 6 compares SCIGAN's and MLP's abilities of predicting overhauls and failures, as well as their ability of prescribing good PM frequencies, for varying levels of selection bias. SCIGAN achieves good predictive performance in terms of MISE for the entire range of operating conditions, ranging from no selection bias and randomly assigned PM frequencies ($\lambda = 0$) to realistic levels of selection bias ($\lambda = 30$). Conversely, the MLP, a supervised approach that does not adjust for selection bias, accurately predicts the potential outcomes when the PM frequencies in the training set are randomized ($\lambda = 0$), but results in notably worse predictions compared to SCIGAN when selection bias is present at higher levels of λ. This result implies that it is important to adjust for dependencies between a contract's characteristics and its observed PM frequency when estimating PM effects from observational data. Similarly, SCIGAN is robust towards higher levels of λ and selection bias in terms of decision-making, illustrated by stable values for PE and PCR across different levels of selection bias, whereas MLP results in less accurate and more costly decisions as bias increases.

Observational data on maintenance operations is likely to contain selection bias, because a machine's PM frequency that was observed in the past will not have been assigned randomly, but based on machine characteristics–be it following a technician's expertise or an existing maintenance policy. Our empirical results demonstrate the importance of dealing with selection bias when working with observational data. Moreover, our results indicate that the generative model in SCIGAN is able to accurately generate counterfactual outcomes to overcome selection bias, resulting in both better predictions and decisions compared to the MLP that does not use this generative model.

6. Conclusion

This work proposes a novel way to optimize the preventive maintenance frequency. Our causal inference approach predicts how the failure and overhaul rate would be impacted by a certain PM frequency, taking the asset's characteristics into account. This is achieved by relying on state-of-the-art machine learning methodologies for causal inference that learn an asset's outcomes for different PM frequencies from observational data on assets that were maintained in the past. The benefit of our approach is that, unlike existing approaches, our methodology does need strong assumptions regarding the failure or overhaul rate or PM effect. These are usually assumed to be known and are difficult to verify from data due to censoring, as assets are usually maintained before failure occurs. Moreover, existing approaches typically do not account for asset heterogeneity. Conversely, our approach is to learn an asset's overhaul and failure rate resulting from a given PM frequency from observational data using flexible machine learning models. This allows to estimate what will happen for a contract given a certain PM frequency, in terms of overhauls and failures, and makes it possible to prescribe the PM frequency that minimizes the combined costs resulting from overhauls, failures, and PM.

Theoretically, we contribute by framing time-based maintenance as a problem of causal inference and by proposing a predict-then-optimize framework to solve this problem. Empirically, we validate our approach with semi-synthetic experiments using real-life data on more than 4,000 full-service maintenance contracts. We find that our proposed approach outperforms both an approach that does not account for selection bias and a non-individualized approach in terms of both accuracy and cost of the prescribed PM schedules. Moreover, we highlight the importance of dealing with selection bias when learning from observational data. Past maintenance decisions were likely not made at random, but based on the asset's characteristics. Because of this, machine learning models need to account for dependencies between

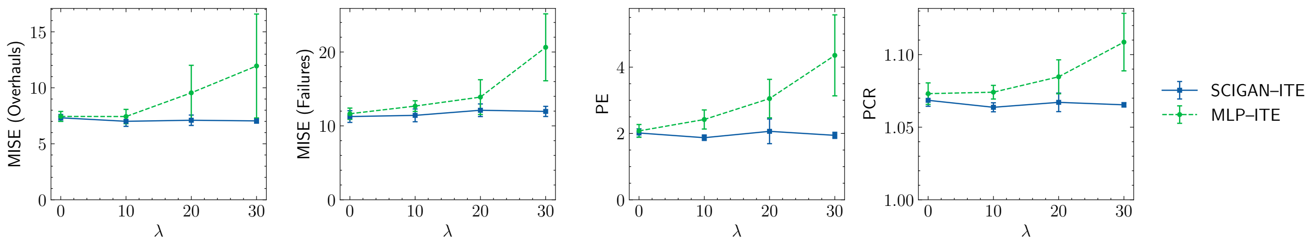


Figure 6: Results for varying levels of selection bias. We show results for different levels of selection bias in terms of λ (see Equation (7)). Although SCIGAN–ITE performs similar to MLP–ITE for lower values of λ, it has better performance for stronger levels of bias in terms of MISE, PE, and PCR.

asset characteristics and the observed PM frequency in order to obtain a good estimate of the overhaul and failure rate for a given PM frequency. These findings show that our proposed approach offers a powerful and flexible policy for individualized maintenance.

6.1. Limitations

Our data-driven approach requires observational data to train the machine learning models. When limited data is available, more simple machine learning methodologies based on, for example, linear regression can be preferred to the presented approach based on neural networks. Choosing and validating causal inference models is an active area of research (Alaa and Van Der Schaar, 2019; Parikh et al., 2022).

Causal inference not only requires data, but also requires that certain assumptions regarding the data are met. The first, overlap, implies that each asset could in principle receive each possible PM frequency, albeit not with the same probability. This requires a degree of flexibility or variability in how PM frequencies were assigned in the past. Alternatively, it might require some experimentation to provide insight into deviations from the existing policy. Overlap can be tested (Lei et al., 2021) and characterized (Oberst et al., 2020) from data. Moreover, recent work has looked at characterizing uncertainty in regions where overlap is violated (Nethery et al., 2019; Jesson et al., 2020).

The second assumption, unconfoundedness, is untestable in practice (Imbens, 2000). However, it can be assessed by people with domain-knowledge that were in charge of making maintenance decisions. The relevant question is whether all relevant information regarding the assignment of past PM frequencies is included in the data. If there are unobserved confounders, adequately adjusting for selection bias might not be possible, which would result in biased estimates of the overhaul and failure rate. Recently, sensitivity analyses have been suggested to assess the influence of hidden confounders (D’Amour, 2019; Franks et al., 2019). Similarly, methods have been proposed for quantifying ignorance regarding the potential outcomes due to possible violations of these assumptions (Jesson et al., 2021).

6.2. Managerial implications

Optimizing maintenance using causal inference and machine learning offers a potentially flexible and powerful maintenance policy. Our approach prescribes the optimal PM frequency to each individual asset by comparing different counterfactual outcomes that would result from different maintenance frequencies, by learning a causal machine learning model from data on assets that were maintained in the past. Under the right conditions, causal inference represents a viable and performant paradigm for maintenance optimization. However, our approach also requires a different way of thinking about maintenance optimization. A completely data-driven policy for preventive maintenance is based on assumptions regarding the data and the models learned from this data. Therefore, maintenance practitioners should check whether their setting allows for causal inference, i.e., whether the requirements presented in Section 4.1 are met. If not, practitioners might consider altering their maintenance operations to satisfy these conditions, e.g., by running small-scale experiments to observe the effect of deviating from their existing policies.

6.3. Future work

In terms of future work, it would be valuable to not only optimize the frequency of one type of PM intervention, but also consider different possible interventions in terms of their depth and costs. This way, it would be possible to alternate cheap, quick visits and more expensive and thorough visits throughout the asset’s lifetime. Moreover, it would be interesting to incorporate more flexible timing of maintenance interventions and consider sequences of different maintenance interventions, potentially prescribed based on real-time dynamic data obtained through sensors. Sequences of treatments have also received attention in the literature on causal inference (e.g., Robins, 1999; Hernán et al., 2001; Bica et al., 2019). Finally, it would be interesting to look at ways of more closely integrating the predictive model in the decision-making step, e.g., by using approaches for integrated predict-and-optimize (Elmachtoub and Grigas, 2022) or cost-sensitive learning (Vanderschueren et al., 2022).

Appendix A Hyperparameter optimization

To make our work more transparent and facilitate the application of our approach, we provide more information regarding the training and hyperparameter optimization of the neural networks used in this work. Table 5 shows training

Name	Range
General	
Batch size	$[32, 64]$
Optimizer	Adam
Learning rate	0.001
GAN	
Hidden neurons	$[16, 32]$
Dosage samples	2
Training iterations	$50,000$
MLP	
Hidden neurons	$[32, 64]$
Training iterations	$10,000$

Table 5
Model training. We show the training settings and hyperparameter ranges that were searched, differentiating between general, GAN-related and MLP-related hyperparameters.

settings and ranges for the different hyperparameters that were searched over, differentiating between general hyperparameters, hyperparameters for the GAN, and hyperparameters for the MLP. For the MLP and MLP–ITE benchmarks, only the general and MLP hyperparameters were searched over. For all models, hyperparameter optimization was done using grid search based on the mean squared error on the observed outcomes in the validation set. For more details regarding SCIGAN's training and optimization, we refer to Bica et al. (2020) and the accompanying repository available at https://github.com/ioanabica/SCIGAN.

References

Ahmad, R., Kamaruddin, S., 2012. An overview of time-based and condition-based maintenance in industrial application. Computers & industrial engineering 63, 135–149.

Alaa, A., Van Der Schaar, M., 2019. Validating causal inference models via influence functions, in: International Conference on Machine Learning, PMLR. pp. 191–201.

Alaswad, S., Xiang, Y., 2017. A review on condition-based maintenance optimization models for stochastically deteriorating system. Reliability engineering & system safety 157, 54–63.

Alves, F., Badikyan, H., Moreira, H.A., Azevedo, J., Moreira, P.M., Romero, L., Leitão, P., 2020. Deployment of a smart and predictive maintenance system in an industrial case study, in: 2020 IEEE 29th International Symposium on Industrial Electronics (ISIE), IEEE. pp. 493–498.

Ansari, F., Glawar, R., Nemeth, T., 2019. Prima: a prescriptive maintenance model for cyber-physical production systems. International Journal of Computer Integrated Manufacturing 32, 482–503.

Athey, S., Wager, S., 2021. Policy learning with observational data. Econometrica 89, 133–161.

Barlow, R., Hunter, L., 1960. Optimum preventive maintenance policies. Operations research 8, 90–100.

Berrevoets, J., Jordon, J., Bica, I., van der Schaar, M., et al., 2020. Organite: Optimal transplant donor organ offering using an individual treatment effect. Advances in Neural Information Processing Systems 33.

Bertsimas, D., Dunn, J., Mundru, N., 2019. Optimal prescriptive trees. INFORMS Journal on Optimization 1, 164–183.

Bertsimas, D., Kallus, N., 2020. From predictive to prescriptive analytics. Management Science 66, 1025–1044.

Bey-Temsamani, A., Engels, M., Motten, A., Vandenplas, S., Ompusunggu, A.P., 2009. A practical approach to combine data mining and prognostics for improved predictive maintenance. Data Min. Case Stud 36.

Bica, I., Alaa, A.M., Jordon, J., van der Schaar, M., 2019. Estimating counterfactual treatment outcomes over time through adversarially balanced representations, in: International Conference on Learning Representations.

Bica, I., Jordon, J., van der Schaar, M., 2020. Estimating the effects of continuous-valued interventions using generative adversarial networks. Advances in Neural Information Processing Systems 33, 16434–16445.

Block, H.W., Borges, W.S., Savits, T.H., 1985. Age-dependent minimal repair. Journal of Applied Probability 22, 370–385.

Bousdekis, A., Lepenioti, K., Apostolou, D., Mentzas, G., 2021. A review of data-driven decision-making methods for industry 4.0 maintenance applications. Electronics 10, 828.

Brown, M., Proschan, F., 1983. Imperfect repair. Journal of Applied Probability 20, 851–859.

Carvalho, T.P., Soares, F.A., Vita, R., Francisco, R.d.P., Basto, J.P., Alcalá, S.G., 2019. A systematic literature review of machine learning methods applied to predictive maintenance. Computers & Industrial Engineering 137, 106024.

Chen, K., Pashami, S., Fan, Y., Nowaczyk, S., 2019a. Predicting air compressor failures using long short term memory networks, in: EPIA Conference on Artificial Intelligence, Springer. pp. 596–609.

Chen, Z., Gryllias, K., Li, W., 2019b. Mechanical fault diagnosis using convolutional neural networks and extreme learning machine. Mechanical systems and signal processing 133, 106272.

Chen, Z., Mauricio, A., Li, W., Gryllias, K., 2020. A deep learning method for bearing fault diagnosis based on cyclic spectral coherence and convolutional neural networks. Mechanical Systems and Signal Processing 140, 106683.

Chukova, S., Arnold, R., Wang, D.Q., 2004. Warranty analysis: An approach to modeling imperfect repairs. International journal of production economics 89, 57–68.

D'Amour, A., 2019. On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative, in: Chaudhuri, K., Sugiyama, M. (Eds.), Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, PMLR. pp. 3478–3486. URL: https://proceedings.mlr.press/v89/d-amour19a.html.

Deprez, L., Antonio, K., Boute, R., 2021. Pricing service maintenance contracts using predictive analytics. European Journal of Operational Research 290, 530–545.

Devriendt, F., Moldovan, D., Verbeke, W., 2018. A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. Big Data 6, 13–41.

Ding, S.H., Kamaruddin, S., 2015. Maintenance policy optimization—literature review and directions. The International Journal of Advanced Manufacturing Technology 76, 1263–1283.

Do, P., Voisin, A., Levrat, E., Iung, B., 2015. A proactive condition-based maintenance strategy with both perfect and imperfect maintenance actions. Reliability Engineering & System Safety 133, 22–32.

Elmachtoub, A.N., Grigas, P., 2022. Smart "predict, then optimize". Management Science 68, 9–26.

Faccio, M., Persona, A., Sgarbossa, F., Zanin, G., 2014. Industrial maintenance policy development: A quantitative framework. International Journal of Production Economics 147, 85–93.

Fast, M., Assadi, M., De, S., 2008. Condition based maintenance of gas turbines using simulation data and artificial neural network: a demonstration of feasibility, in: Turbo Expo: Power for Land, Sea, and Air, pp. 153–161.

Figueroa Barraza, J., Guarda Bräuning, L., Benites Perez, R., Morais, C.B., Martins, M.R., Droguett, E.L., 2022. Deep learning health state prognostics of physical assets in the oil and gas industry. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability 236, 598–616.

Fouladirad, M., Paroissin, C., Grall, A., 2018. Sensitivity of optimal replacement policies to lifetime parameter estimates. European Journal of Operational Research 266, 963–975.

Franks, A., D'Amour, A., Feller, A., 2019. Flexible sensitivity analysis for observational studies without observable implications. Journal of the American Statistical Association .

Gits, C., 1992. Design of maintenance concepts. International journal of production economics 24, 217–226.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Communications of the ACM 63, 139–144.

Hernán, M.A., Brumback, B., Robins, J.M., 2001. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. Journal of the American Statistical Association 96, 440–448.

Hirano, K., Imbens, G.W., 2004. The propensity score with continuous treatments. Applied Bayesian modeling and causal inference from incomplete-data perspectives 226164, 73–84.

Holland, P.W., 1986. Statistics and causal inference. Journal of the American Statistical Association 81, 945–960.

Imai, K., Van Dyk, D.A., 2004. Causal inference with general treatment regimes: Generalizing the propensity score. Journal of the American Statistical Association 99, 854–866.

Imbens, G.W., 2000. The role of the propensity score in estimating dose-response functions. Biometrika 87, 706–710.

Jansen, F., Holenderski, M., Ozcelebi, T., Dam, P., Tijsma, B., 2018. Predicting machine failures from industrial time series data, in: 2018 5th International Conference on Control, Decision and Information Technologies (CoDIT), IEEE. pp. 1091–1096.

Jesson, A., Mindermann, S., Gal, Y., Shalit, U., 2021. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding, in: International Conference on Machine Learning, PMLR. pp. 4829–4838.

Jesson, A., Mindermann, S., Shalit, U., Gal, Y., 2020. Identifying causal-effect inference failure with uncertainty-aware models. Advances in Neural Information Processing Systems 33, 11637–11649.

de Jonge, B., Klingenberg, W., Teunter, R., Tinga, T., 2015. Optimum maintenance strategy under uncertainty in the lifetime distribution. Reliability engineering & system safety 133, 59–67.

de Jonge, B., Scarf, P.A., 2020. A review on maintenance optimization. European Journal of Operational Research 285, 805–824.

Kijima, M., 1989. Some results for repairable systems with general repair. Journal of Applied Probability 26, 89–102.

Kusiak, A., Verma, A., 2011. A data-mining approach to monitoring wind turbines. IEEE Transactions on Sustainable Energy 3, 150–157.

Lee, Y.L., Juan, D.C., Tseng, X.A., Chen, Y.T., Chang, S.C., 2017. Dc-prophet: Predicting catastrophic machine failures in datacenters, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer. pp. 64–76.

Lei, L., D'Amour, A., Ding, P., Feller, A., Sekhon, J., 2021. Distribution-free assessment of population overlap in observational studies .

Leukel, J., González, J., Riekert, M., 2021. Adoption of machine learning technology for failure prediction in industrial maintenance: A systematic review. Journal of Manufacturing Systems 61, 87–96.

Liu, Y., Huang, H.Z., Zhang, X., 2011. A data-driven approach to selecting imperfect maintenance models. IEEE Transactions on Reliability 61, 101–112.

Louit, D.M., Pascual, R., Jardine, A.K., 2009. A practical procedure for the selection of time-to-failure models based on the assessment of trends in maintenance data. Reliability Engineering & System Safety 94, 1618–1628.

Lu, Y., Sun, L., Zhang, X., Feng, F., Kang, J., Fu, G., 2018. Condition based maintenance optimization for offshore wind turbine considering opportunities based on neural network approach. Applied Ocean Research 74, 69–79.

Malik, M.A.K., 1979. Reliable preventive maintenance scheduling. AIIE Transactions 11, 221–228.

Matyas, K., Nemeth, T., Kovacs, K., Glawar, R., 2017. A procedural approach for realizing prescriptive maintenance planning in manufacturing industries. CIRP Annals 66, 461–464.

Nakagawa, T., 1979a. Imperfect preventive-maintenance. IEEE Transactions on Reliability 28, 402–402.

Nakagawa, T., 1979b. Optimum policies when preventive maintenance is imperfect. IEEE Transactions on Reliability 28, 331–332.

Nemeth, T., Ansari, F., Sihn, W., Haslhofer, B., Schindler, A., 2018. Prima-x: A reference model for realizing prescriptive maintenance and assessing its maturity enhanced by machine learning. Procedia CIRP 72, 1039–1044.

Nethery, R.C., Mealli, F., Dominici, F., 2019. Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality. The annals of applied statistics 13, 1242.

Oberst, M., Johansson, F., Wei, D., Gao, T., Brat, G., Sontag, D., Varshney, K., 2020. Characterization of overlap in observational studies, in: International Conference on Artificial Intelligence and Statistics, PMLR. pp. 788–798.

Orrù, P.F., Zoccheddu, A., Sassu, L., Mattia, C., Cozza, R., Arena, S., 2020. Machine learning approach using mlp and svm algorithms for the fault prediction of a centrifugal pump in the oil and gas industry. Sustainability 12, 4776.

Parikh, H., Varjao, C., Xu, L., Tchetgen, E.T., 2022. Validating causal inference methods, in: International Conference on Machine Learning, PMLR. pp. 17346–17358.

Pham, H., Wang, H., 1996. Imperfect maintenance. European Journal of Operational Research 94, 425–438.

Poppe, J., Boute, R.N., Lambrecht, M.R., 2018. A hybrid condition-based maintenance policy for continuously monitored components with two degradation thresholds. European Journal of Operational Research 268, 515–532.

Robins, J.M., 1999. Association, causation, and marginal structural models. Synthese , 151–179.

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of educational Psychology 66, 688.

Rubin, D.B., 2004. Direct and indirect causal effects via potential outcomes. Scandinavian Journal of Statistics 31, 161–170.

Rubin, D.B., 2005. Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association 100, 322–331.

Savitha, R., Ambikapathi, A., Rajaraman, K., 2020. Online rbm: Growing restricted boltzmann machine on the fly for unsupervised representation. Applied Soft Computing 92, 106278.

Schwab, P., Linhardt, L., Bauer, S., Buhmann, J.M., Karlen, W., 2020. Learning counterfactual representations for estimating individual dose-response curves, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 5612–5619.

Silva, R., 2016. Observational-interventional priors for dose-response learning. Advances in Neural Information Processing Systems 29.

Swanson, L., 2001. Linking maintenance strategies to performance. International journal of production economics 70, 237–244.

Tanwar, M., Rai, R.N., Bolia, N., 2014. Imperfect repair modeling using kijima type generalized renewal process. Reliability Engineering & System Safety 124, 24–31.

Tian, Z., 2012. An artificial neural network method for remaining useful life prediction of equipment subject to condition monitoring. Journal of intelligent Manufacturing 23, 227–237.

de Toledo, M.L.G., Freitas, M.A., Colosimo, E.A., Gilardoni, G.L., 2015. Ara and ari imperfect repair models: Estimation, goodness-of-fit and reliability prediction. Reliability Engineering & System Safety 140, 107–115.

Vanderschueren, T., Verdonck, T., Baesens, B., Verbeke, W., 2022. Predict-then-optimize or predict-and-optimize? an empirical evaluation of cost-sensitive learning strategies. Information Sciences 594, 400–415.

Varian, H.R., 2016. Causal inference in economics and marketing. Proceedings of the National Academy of Sciences 113, 7310–7315.

Verbeke, W., Olaya, D., Berrevoets, J., Verboven, S., Maldonado, S., 2020. The foundations of cost-sensitive causal classification. arXiv preprint arXiv:2007.12582 .

Verbeke, W., Olaya, D., Guerry, M.A., Van Belle, J., 2022. To do or not to do? cost-sensitive causal classification with individual treatment effect estimates. European Journal of Operational Research .

Wang, H., 2002. A survey of maintenance policies of deteriorating systems. European Journal of Operational Research 139, 469–489.

Webbink, D., 2005. Causal effects in education. Journal of Economic Surveys 19, 535–560.

Wu, B., Tian, Z., Chen, M., 2013. Condition-based maintenance optimization using neural network-based health condition prediction. Quality and Reliability Engineering International 29, 1151–1163.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., Zhang, A., 2021. A survey on causal inference. ACM Transactions on Knowledge Discovery from Data (TKDD) 15, 1–46.

Ye, Z., Yu, J., 2021. Aksnet: A novel convolutional neural network with adaptive kernel width and sparse regularization for machinery fault diagnosis. Journal of Manufacturing Systems 59, 467–480.

Zhang, M., Xie, M., 2017. An ameliorated improvement factor model for imperfect maintenance and its goodness of fit. Technometrics 59, 237–246.

Zhao, J., Huang, W., 2021. Transfer learning method for rolling bearing fault diagnosis under different working conditions based on cyclegan. Measurement Science and Technology 33, 025003.