FORECASTING CASH FLOWS AT AN INDEPENDENT TREATMENT CENTRE

MSc Business Administration – Financial Management

Faculty of Behavioural, Management and Social Sciences (BMS)

Anouk Rolink
S1929844

Supervisors
Dr. R.A.M.G. Joosten (University of Twente)
Dr. P.C. Schuur (University of Twente)
Dr. A.G.J. Bot MBA (OCON)

August, 2023

Preface

Dear reader,

During my bachelor's in Health Sciences, two fields of science specifically raised my interest and enthusiasm, being operations- and financial management. For me, the obvious next step was to follow both masters Health Sciences and Business Administration (BA) – Financial Management at the University of Twente. For the former, I finished my thesis at OCON Orthopedische Kliniek. Because of their focus on continuous improvements, engagement in scientific research, and pleasant working environment, I was glad to perform my second master thesis at OCON too.

The result of working at OCON from February until August 2023 lies before you, my master thesis Business Administration called 'Forecasting cash flow at an independent treatment centre'. Because of a change of the Dutch healthcare funding system, hospitals face a financial management challenge. They lack insight into the timing and amount of future cash flows resulting from current care activities. That is why the goal of this study was to develop a forecasting model for the operating cash flows at OCON.

I would like to thank Reinoud Joosten and Peter Schuur, my supervisors from the UT, for their input, constructive feedback, and support in supervising this research. I would also like to thank Arjan Bot, my supervisor at OCON. Without his enthusiastic guidance and extensive knowledge of OCON's financial management, I could not have written this master thesis. Thanks also go to all other employees and fellow students at OCON. Not only did they support me over the course of my studies, but they also made my period of well over a year at OCON very enjoyable. As a result, I can proudly present this master thesis to you.

I hope you will enjoy reading my master thesis!

Anouk Rolink

Hengelo, August 2023

Management summary

Introduction

With the goal of achieving high-quality, universally accessible healthcare, a reform of the Dutch healthcare system was introduced in 2006. In this new system, every citizen is ensured of a broad basic package of healthcare by a mandatory insurance, and provision and payment of care is performed by private, competing healthcare insurers and providers. Diagnosis Treatment Combinations (DTCs) are used for claiming the costs and cover the entire care process associated with a specific diagnosis. To improve transparency and uniformity in the pricing of medical services, the model was revised in 2012 by the introduction of the DTCs towards transparency (DOT) registration system. The core idea behind DOT is that the amount of the claim is based on the care that was provided, which can only be assessed after the course of treatment has been finished.

Consequently, hospitals face a financial management challenge. Whereas costs could previously be claimed as soon as a diagnosis was made, they now must wait until the end of the DTC care path before invoicing, which can vary in duration with a maximum of 120 days. This results in both a delay between the provision of care and the corresponding future cash flows, as well as ex-ante uncertainty regarding the timing and magnitude of the financial flows. This is caused by the fact that the official DTC healthcare product and related charges are only established once the DTC care path has been closed. This challenge complexifies cash flow forecasting, and practice shows that hospitals face difficulties to make accurate forecasts. As valid forecasts are essential to the effective running of operations, to informed strategic decision making, investments, and to firm survival, hospitals explore methods to improve forecast reliability.

Case study

One of these hospitals is OCON Orthopedische Kliniek (OCON), a financially autonomous hospital specialised in orthopaedic and sports medical care, with departments in the hospitals of Ziekenhuis Groep Twente (ZGT) in both Hengelo and Almelo. In their aim for continuous improvements, OCON wishes to gain more accurate forecasts of the cash inflow generated by the provision of care. This would result in better financial management information and allows for adjusting business processes accordingly. That is why we aim at developing a model for forecasting the cash flows resulting from providing health care services. Therefore, the central research question is:

"How can OCON's cash flow from operations be forecasted?

Literature review

Following a literature search, we identified only one study with a similar problem context and research goal as our study. Using classifier and process prediction algorithms, the authors predicted the sequence of treatments administered, the duration of the care path and the final billable care product for three specific diagnoses with an average forecast accuracy of 47%. No other studies concerning cash flow forecasts from operations in hospitals were identified in the literature.

Literature regarding cash flow forecasting predominantly focuses on stock-listed companies based in the United States. The two most used predictors for forecasting cash flows are cash flows from operations and earnings, and literature lacks consensus on the superiority of one over the other. Multiple studies have shown that disaggregating earnings into their major accrual components significantly enhances the predictive performance of earnings on future operating cash flows. Accruals include income that a firm has earned but has not yet been paid for, e.g., change in inventory.

Methodology

Inspired by methods of the above-described study for forecasting care path duration and final billable care product, we explored the possibility to follow a similar approach for our study. However, translation to our situation implies the development of at least 15 separate forecasting models, corresponding to the most frequent care products, to predict the overall operating cash flow. This was not feasible for us, which is why we have chosen to apply linear regression models to forecast two outcome measures related to operating cash flows, being work-in-progress (WIP) and revenue. WIP represents the monetary value of all care paths that have not been closed yet.

Based on the above definition of accruals, we consider the monetary value of activities in care paths accruals as the expenses for providing care have been incurred, but the hospital has yet to receive payment. Similarly, a change in production capacity can be regarded as a change in inventory. Therefore, capacity indicators could serve as predictors of WIP and revenue. To decrease standard deviations of forecasts and increase accuracy, we developed separate forecasting models for conservative and operative WIP. We included capacity indicators relating to the outpatient clinic and operating room (OR) to predict conservative and operative WIP, respectively. For forecasting revenue, we did not distinct between conservative or operative parts, and included all variables in one prediction model. In all models, we included dummy variables to correct for COVID-19 and capacity reduction periods as these circumstances complexify accurate cash flow forecasting.

After data preparation, we selected variables using the backward variable selection method. P-values of smaller than 0.05 were considered statistically significant. We assessed goodness of fit by evaluating adjusted R^2 and root-mean-square error (RMSE), and also considered forecast accuracy as performance measure. For forecasting WIP, we used a validation set to evaluate external validity.

Results

WIP datasets had a sample size of 48 and raw revenue data comprised 20 observations. From the raw revenue data comprising 20 observations, we generated 21 additional monthly revenue observations from quartile data, resulting in an adjusted revenue sample size of 41. Because of these small sample sizes, we should draw conclusion with caution as chance considerably impacted our results.

Results show that 71% of the variation in conservative WIP for the next month can be explained by the number of appointments at the outpatient clinic in the current and previous month, when corrected for COVID-19 and reduction. Median forecast accuracy was 96.2%. When applying the model to new data, adjusted R^2 and forecast accuracy decreased to 49% and 95.3%, respectively, indicating limited model generalizability. Similarly, adjusted R^2 was 83% and median forecast accuracy 95.4% when including the variables scheduled OR-hours for the next and current month to forecast operative WIP for the next month. We also corrected for COVID-19 by inclusion of the dummy variable. This model seemed to have better generalizability compared with the former model, as forecasting accuracy did not considerably differ between the training and test dataset and adjusted R^2 decreased with 12%. For both models, all variable estimates were in line with our expectations.

Forecasting revenue on the raw data resulted in and adjusted R^2 of 0.38 and median forecast accuracy of 93.2% when using the number of outpatient clinic appointments three months ago and the scheduled and available OR-hours current and previous month as predictors. For the adjusted data, number of appointment two months ago, available OR-hours previous month, and scheduled OR-hours current month have proven to be the best predictors, resulting in a model with an adjusted R^2 of 0.25 and mean forecast accuracy of 93.1%. In this model, we also included the dummy that corrects for a reduction period four months ago. Both revenue forecasting models comprised intuitive and counterintuitive variable estimates.

Conclusion

The models forecasting WIP outperform the ones predicting revenue based on goodness of fit and forecast accuracy. However, because of the small sample sizes, we currently do not recommend applying any of the forecasting models in practice, but rather use them as a blueprint for the future when more, good quality data is available. Adding these data are expected to enhance model accuracy, reliability, and generalizability.

Despite the lack of statistical power, our results still provide an estimation of the impact various capacity variables have on financial outcomes and their magnitude and timing. Therefore, our results concretised the expected relationship between capacity and financial management at OCON.

Contents

Section 1

Introduction

A reform of the healthcare system was introduced in January 2006, to achieve high-quality, universally accessible healthcare in the Netherlands (Boot, 2013; Tweede Kamer der Staten-Generaal, 2004). Based upon solidarity and affordability, this system initiated regulated competition between healthcare provider and between healthcare insurers, and introduced the Health Insurance Act (*Zorgverzekeringswet).* This law ensures that every Dutch citizen has access to a broad basic package of health care by a mandatory insurance, and it transfers provision and payment of care to private, competing health care insurers and providers. As a result, three markets have emerged in the Dutch healthcare system, which is schematically shown in Figure 1.

At the insurance market, the insurance-obliged consumer pays a monthly premium to the acceptance-obliged health care insurer (Boot, 2013; Tweede Kamer der Staten-Generaal, 2004). In return, the insurer reimburses any costs, aside from the deductible, related to health care from the basic package. To trigger competition, consumers are given free choice of insurer. At the purchasing market, insurers negotiate with health care providers about the quality, quantity, and costs of health care services before purchasing care on behalf of their insured. The competition for contracts with insurers would serve as an incentive for healthcare providers to offer high-quality care at the lowest possible cost. Providing care takes place at the delivery market and does not involve financial interactions between patient and health care professional. Competition between healthcare providers is stimulated as patients are free to choose the healthcare provider of their preference, as long as their insurer has entered into a contract with the healthcare provider. In short, healthcare providers submit their patients' treatment bills to the patient's insurer, who subsequently bill their customer.

Claiming the costs for the provided care in the new health care system was performed using Diagnosis Treatment Combinations (DTCs). Each DTC covers the entire care process associated with a specific diagnosis, from the initial appointment with the medical specialist through the end of the course of treatment (Boot, 2013; Minister of Health Welfare and Sport J.F. Hoogervorst, 2003).

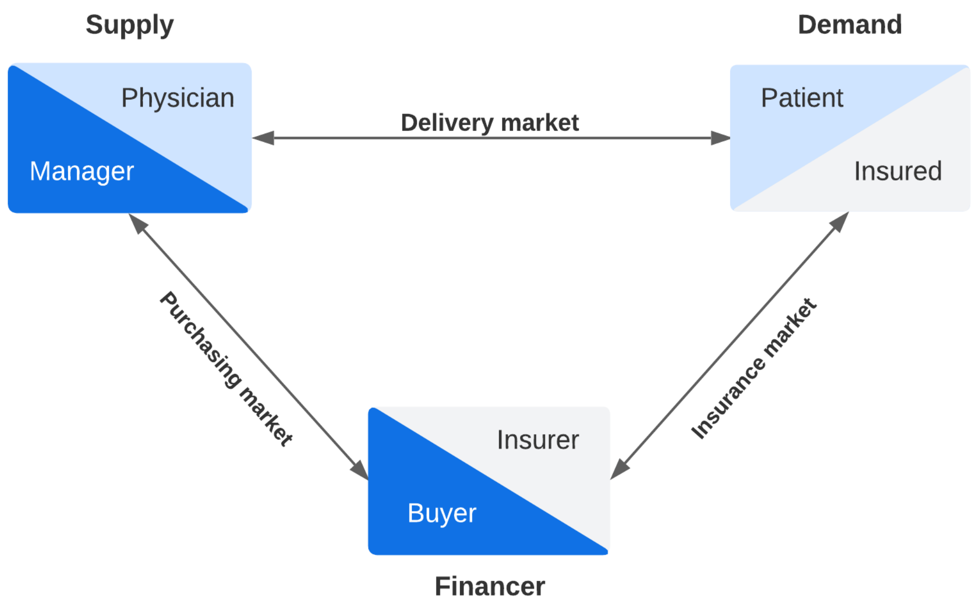

Figure 1. Schematic representation of the three markets in the Dutch healthcare system (Boot, 2013).

The amount of money to be claimed was established by the diagnosis for which typical costs would be agreed upon by hospitals and insurance providers. However, after implementation this funding system had proven to be complex and inefficient. Moreover, it was unclear for patients how much their care had cost. A revision of the funding model was introduced in January 2012 to provide more transparency and uniformity in the pricing of medical services: the DOT registration system (Jeurissen & Maarse, 2021). DOT stands for "DTCs towards transparency" in Dutch. The fundamental concept of DOT is that the amount of the claim is determined by the actual care rendered, which can only be determined after the course of treatment has been completed.

As a result of this development, hospitals now face a challenge in their financial management. They have to wait until the end of the DTC care path before claiming costs, which can vary in duration with a maximum of 120 days, whereas they previously could invoice as soon as the diagnosis was made (Nederlandse Zorgautoriteit, 2022). Besides the fact that this generates a delay between the provision of care and the associated cash inflows for the hospital, it also creates ex-ante uncertainty regarding the timing and amount of the financial flows. This is because the official DTC healthcare product and associated costs are only determined when the DTC care path closes.

As applies to any business, accurate cash flow forecasts are essential to the effective running of hospitals, as cash is vital to firm survival, informed strategic decision making, and making investments (McLaney & Atrill, 2018). Cash flow predictions are also of interest for healthcare providers in honouring the agreements they made with insurers regarding the maximum quantities of care to be delivered. Because of the delay in rendering care and invoicing, healthcare providers only know how many patients with specific diagnoses they have treated when it is already a fait accompli. This could even result in the provision of free care if the healthcare provider exceeds the turnover limit, meaning that they have treated more patients than was agreed upon during the contract negotiations with insurers.

Practice shows that hospitals face difficulties to predict their cash flows. That is why we aim at developing a model for forecasting the cash flows by performing a case study at OCON. We focus on cash inflows from operations, which we define as the amount of money earned by providing health care services. With the goal of providing high-quality, tailor-made care, OCON was established in 2010 as a hospital specialized in orthopaedic and sports medical care, located in the hospitals of Ziekenhuis Groep Twente in Hengelo and Almelo. Currently, OCON's workforce comprises 19 orthopaedic surgeons, 5 sports medicine physicians, 6 anaesthesiologists, and around 250 employees that annually treat 25,000 unique patients. Approximately 20 percent of these patients undergo surgery. Autonomy was acquired in 2019. Hence, annual contract negotiations with insurers take place to determine the quality, quantity, and price of healthcare in the upcoming year. In their aim for continuous improvements, OCON wishes to gain more accurate forecasts of the cash inflows generated by the provision of care. This would result in better financial management information and allows for adjusting business processes accordingly.

Therefore, the central research question is:

"How can OCON's cash flow from operations be forecasted?

This paper consists of 6 sections, including this introduction. In this first section, we have provided a basic introduction to the organisation of the health care system in the Netherlands and the problem context of this study. In Section 2, we elaborate on the DOT registration system for the billing of specialised medical care. In Section 3, we review the available literature concerning cash flow forecasting, particulartly focused on the healthcare sector. Subsequently, we describe the research methods used to answer the central research question in Section 4. In Section 5 we present the results. Finally, we discuss the results and give an answer to the central research question in Section 6. In this final section, we also discuss the managerial implications, the limitations and strengths, and the suggestions for further research.

Section 2

DOT registration system

In this section, we explain the DOT registration system for the financing of specialist medical care in the Netherlands in more detail. All information in this section is derived from the manual DOT registration system and the regulation specialist medical care as published by the Dutch Healthcare Authority (Nederlandse Zorgautoriteit, 2022, 2023).

2.1 Definitions

To facilitate explanation, let us first provide some definitions in this section.

DTC care product. As mentioned in Section 1, a DTC comprises all actions that are needed to diagnose and treat a patient with a specific diagnosis. A DTC of a specific diagnosis is also called a care product. The price of a care product is an average of all costs associated with activities corresponding to that specific diagnosis, which means that not every component is billed separately. The DOT registration system contains approximately 4,400 unique DTC care products.

DTC care path. A DTC care product consists of multiple care paths. A path is a demarcated period of maximum 120 days within the DTC product for which the provided care is invoiced once the path is closed. The duration and rules for closing a path are described in the registration rules, which we will discuss below in more detail.

Product structure. Product structure relates to the way in which the diagnosis and treatment actions are translated to a DTC product. This is based on the global standard for classification of diagnoses called the International Statistical Classification of Diseases and Related Health Problems (ICD-10) (World Health Organisation, 2023).

Grouper. An automated system which derives the DTC care product from the care activities in the care path based on the product structure and associated decision rules.

2.2 The four phases of the DOT registration system

The DOT registration system is based on the RSDI-model: Register, Summarise, Derive, and Invoice. We describe these four phases separately in the following sections. Because of the extensiveness and complexity of the system, the descriptions apply to the vast majority of OCON's patients and do not elaborate on exceptional cases, such as patients admitted to the intensive care.

2.2.1 Register

All healthcare activities related to the diagnosis and treatment of a patient's care need are registered in the electronic patient record in the hospital's digital environment. At the first consultation with a medical specialist, the DTC product and automatically the initial care path, are opened. Besides the registration of all activities, the treating medical specialist registers a diagnosis. At the opening of the care path, this is a provisional diagnosis. The medical specialist can adjust the diagnosis throughout the duration of the path to make it best suited to the care needed. A final typical diagnosis for the path is determined when the path closes.

For each patient, only one DTC product per specific care need can be opened at a medical specialty. This means that it is not allowed to open a new product if another one for that particular care need is still active. For example, if a patient has an opened care product for left knee pain due to osteoarthritis, all activities related to the treatment of the knee osteoarthritis are registered in the already opened product, instead of opening a new one in case of additional treatments.

2.2.2 Summarise

When a care path is closed, all registered information is summarised in a dataset. This dataset is used in the Derive phase to determine the amount of money that can be invoiced. Closing and summarising of care paths is automated as the closing rules are integrated in the hospital's digital environment. However, to support deeper understanding of the problem context, the closing rules are outlined below.

Multiple types of care paths exist, but here we only focus on initial- and follow-up ones. Figure 2 shows three possible DTC products and the paths they comprise. Closure of a care path initiates the Derive and Invoice phases, which eventually results in a bill being paid by the insurer. That is further described in Sections 2.2.3 and 2.2.4.

Closing rules vary between care paths with conservative and surgical treatments. Conservative, initial care paths are closed on the 90th day after opening (Figure 2a). Because care products comprise a maximum of one initial care path, the subsequent care paths are follow-up care paths. These have a duration of 120 days in case of conservative treatment (Figure 2a).

Regardless of being initial or follow-up, care paths with surgical treatments are closed on the 42nd day after discharge from the hospital. Figure 2b shows a care path in which a surgery was performed on the 10th day of the first follow-up care path, which is why the care path has closed after 52 (10 + 42) days. As care paths have a maximum duration of 120 days, some surgical care paths close earlier than 42 days after discharge if the term of 120 days is reached. This is shown in Figure 2c, where surgery took place on the 90th day of the first follow-up care path. As 132 (90 + 42) days exceeds the maximum of 120 days, the care path is closed after 120 days instead of 132.

The entire DTC care product closes after three subsequent empty care paths, being a period of 3*120 days without registered health care activities and without scheduled appointments in the future.

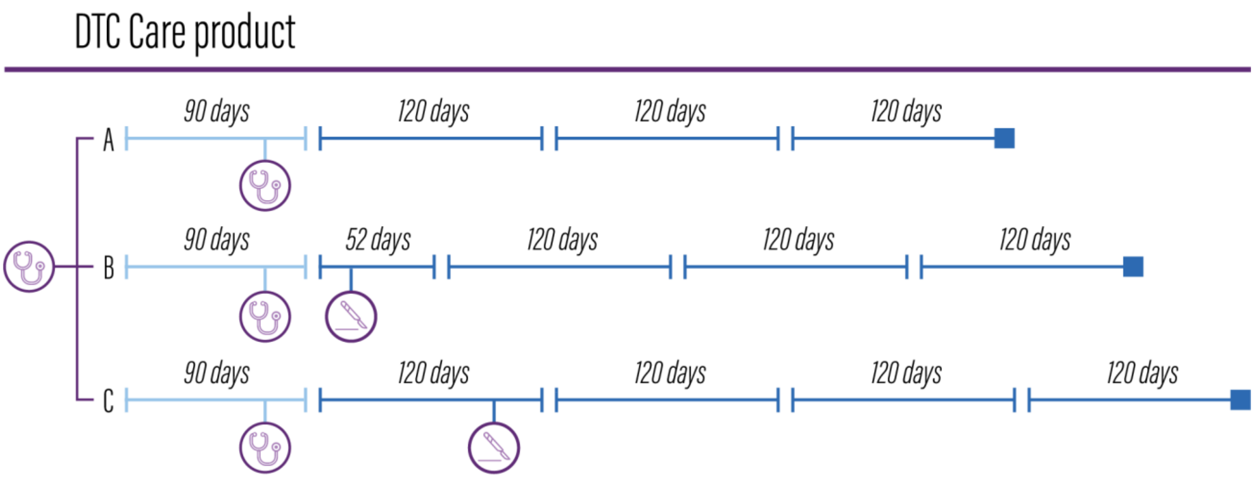

Figure 2. Visual presentation of possible DTC Care products. A stethoscope icon represents a conservative appointment at the outpatient clinic and a scalpel icon represents a surgery. Initial and follow-up care paths are coloured light blue and dark blue, respectively. The stethoscope icon on the left indicates the opening of the initial care paths at the first consultation with a medical specialist. Every vertical line at the end of a care path indicates its closure, resulting in the automatic opening of a subsequent care path the next day. The square represents the closing of the DTC care product.

2.2.3 Derive

The summarised dataset, also called declaration dataset, is sent to the grouper. At OCON, this happens once a week, on Tuesdays. Before deriving a billable care product, the grouper performs some checks, for example to verify whether all necessary information is available, such as a registered typical diagnosis and the validity of the registered care activities. If all requirements are fulfilled, the grouper translates the declaration dataset into a result set based on the product structure and associated decision rules.

After the grouper has extracted care products from the declaration dataset, the result set is returned to the healthcare provider. This result set contains care product- and declaration codes. Additionally, hash codes, functioning as a seal, are included to guarantee that a grouper run has taken place and no changes have been made between receiving the result set and invoicing the insurer.

It should be noted that not every possible combination of care activities and typical diagnosis is deduced to a unique billable care product. Every care product is based on an average of the activities needed for the treatment of a specific care need, and cut-off points for decision making have been identified to facilitate derivation. This means that there is room for differences in care activities between patients with the same care products. To clarify, imagine a situation of three patients whose care path has just closed with a typical diagnosis of hip osteoarthritis. During their care paths, the first patient had one appointment at the outpatient clinic, the second one had two appointments and the third one had three. Based on the decision rules, the grouper will derive the same care product for patients one and two, whereas the care path of patient three is categorised as another care product. This is explained by the fact that a maximum of two outpatient clinic appointments is the cut-off point. One or two appointments results in the same care product, but more than two results in another, more expensive care product.

Registration errors can result in an illogical combination of care activities and typical diagnoses, which leads to a faulty, non-claimable care product by the grouper. The grouper returns this specific care product to the healthcare provider who is responsible for correction before resending it to the grouper. Only if a care product has successfully been processed by the grouper, an invoice can be sent. An example of an incorrect declaration dataset is the registration of a hip prothesis for a patient with a diagnosis related to the knee. This might occur for patients with parallel care products for separate care needs and diagnoses, such as a patient with simultaneous treatment processes for knee and hip osteoarthritis.

2.2.4 Invoice

After closure of a care path and successful translation of the declaration dataset to the result set, the healthcare provider can settle the costs with the healthcare insurers. In the hospital's digital environment, the healthcare products are matched with their corresponding prices. These are the prices resulting from the negotiations with the insurers and are fixed rates throughout the year. After invoicing, which happens on Fridays, bills are usually paid within 30 days. The grouper serves as the standard for invoicing specialist medical care and only functions in case of correct DTC registration, which is why invoices are generally accepted by insurers and are rarely rejected.

2.3 Managerial challenges

As already mentioned in the introduction, the reform of the funding of the Dutch healthcare system to the DOT registration system has resulted in some managerial challenges. In this section, these are explained in more detail to clarify the problem context of this study.

At the basis of these managerial challenges lies the core idea of the DOT system that the amount of a claim is determined by the actual care provided, which can only be assessed after a care path has been closed. This also follows from the RSDI-methodology as described in Section 2.2. Firstly, the typical diagnosis can be altered throughout the duration of the care path, until the diagnosis is final at closure. As the diagnosis is important for the grouper in deriving the care product, this implies that it is uncertain to which care product and corresponding economic value a care path will be translated until the path is closed. Similarly, care activities can be registered until the final day of the care path. These care activities could result in exceeding a cut-off point and thus result in another care product being returned by the grouper. In addition to the uncertainty regarding the amount of future cash flows resulting from current care activities, variability in the timing of future cash flows was also introduced by the DOT system. According to the closing rules, care paths with conservative treatments can last 90 or 120 days. Care paths with surgical treatments on the other hand have increased variability, as these can have a duration of any value between 42 and 120 days.

To facilitate the effective running of OCON, accurate cash flow predictions are vital. Needless to say, the above-mentioned managerial challenges complexify valid cash flow forecasting. Currently, forecasts are based on previous years, but these have proven of insufficient reliability. Causes are found in the limited amount of data from previous years, as autonomy was gained only four years ago, and the impact of the COVID-19 pandemic on the usability of the data from years 2020 and 2021. Additionally, substantial business changes have taken place over the past few years, such as adjustments in the scheduling processes. These changes impact the finances resulting from healthcare activities. Finally, by merely basing forecasts on previous years, the influence of contemporary changes in the business environment are neglected, making predictions inherently less reliable. This way of reactive cash flow management instead of proactive is something which OCON regards as undesirable. However, no better method is available at this moment. By examining how cash flows can be forecasted, this study aims to contribute to enhanced financial management information, thereby the overall financial health of the company. Additionally, this study contributes to filling the academic research gap concerning cash flow forecasts of hospitals in the Netherlands or other healthcare facilities or firms facing challenges regarding the timing and amount of their expected cash flows. We discuss the current literature in Section 3.

Section 3

Theoretical framework

In this section, we explore the current theory regarding cash flow forecasting and predictive models. In Section 3.1, we present literature regarding financial forecasting in hospitals. In Section 3.2, we discuss studies on the prediction of future cash flows in firms, and in Section 3.3 we summarize some literature regarding prediction model techniques.

3.1 Financial forecasting in hospitals

The literature on forecasting cash flows in hospitals, especially in the Netherlands, is sparse. However, van der Spoel et al. (2013) investigated the use of process mining to predict the cash flow of a hospital in the Netherlands. The organisational change of the Dutch healthcare system in 2012 and the resulting financial management challenges served as the motivation for conducting their research. The sequence of the treatments administered, the duration, and the final billable care product were predicted using classifier and process prediction algorithms. The authors used three datasets, all containing data related to a specific cardiology or lung care diagnosis. The random forest algorithm turned out to be the best classifier in predicting the final care product and its cost, with a mean of 55%, 47%, and 40% accuracy for the three investigated diagnoses. Based on the diagnosis, the most recent care activity, and the predicted care product, the remaining activities could be predicted with 80% precision using the Most Likely Path algorithm. The authors concluded that their results can be used to analyse the cash flow for other hospitals with a few modifications. These adjustments are recommended because of the significant influence of noise in their data on the results. Noise is defined as the presence of unrelated activities in the process that is analysed.

No other studies concerning cash flow forecasts from operations in hospitals were identified in the literature. However, some studies in health care facilities have focused on identifying indicators of financial distress or developed prediction models to forecast the risk thereof. Coyne & Singh (2008) identified predictors of financial failure by comparing healthcare providers that filed bankruptcy with solvent ones. Their findings indicate distinct financial trends between the two, particularly for the operating-cash-flow-related measures. Financially stable health care institutions had significantly higher ratios of operating cash flow to previous operating cash flows, to net revenues, and to total liabilities than the bankrupt ones. This indicates the importance of active cash flow management, thereby the importance of adequate cash flow forecasts.

Holmes, Kaufman & Pink (2017) developed and validated a logistic regression model to predict the probability of financial distress and closure within 2 years for rural hospitals in the United States. Results showed that smaller hospitals with lower profitability, less reinvestment, lower market share and poorer economic conditions are more prone to financial distress. Besides the questionable external validity of these results, the Dutch healthcare system drastically differs from the system in the United States, these studies and their used methods are not directly applicable to our research because of the essentially different research goal.

Although solvency analysis is of critical importance for the financial health of hospitals, our study focuses on predicting the operating cash flows rather than identifying indicators of financial distress, which is why solvency ratios and financial distress indexes are not considered appropriate methods for our study. However, using regression analysis as a forecasting model might be an option. This is explored in further detail in the following sections.

3.2 Cash flow forecasting in firms

As mentioned before, future cash flows are essential to a company's survival, making reliable cash flow forecasting crucial. Before the 1980s, accruals and non-cash items were subtracted from earnings to estimate cash flows indirectly, making measurement mistakes unavoidable (Drtina & Largay III, 1985). Since, there has been an increase in the number of papers studying cash flow predictions. The assertion of the Financial Accounting Standards Board implying that earnings based on accrual accounting are better in predicting future operating cash flows than the current operating cash flows themselves raised additional interest.

Therefore, Mulenga & Bhatia (2017) have reviewed the academic literature on predicting firms' future operating cash flow. Roughly, the two most used predictors are identified, being cash flow from operations and earnings. The authors included forty studies that compared the accuracy of both as predictors of future cash flow. Most of these studies used data of firms listed on country-specific stock exchanges and 33 applied simple or multiple regression equations to calculate the respective predictive abilities of operating cash flow and earnings. To assess accuracy, the studies mostly used R-squared values and compared forecast errors between the two models. Approximately a quarter of the studies indicate superiority of operating cash flow as predictor of future cash flows but results also showed that consensus regarding the best predictor of future operating cash flows in firms lacks.

For example, Habib (2010) examined the ability of both predictors to forecast future operating cash flows in Australia, and showed that, although moderated by firm-specific contextual factors, the cash flow-based models outperform the earnings-based models in terms of accuracy. Based on 14-years data of 12263 observations, the earning-based models had an average forecast error of 0.095 in predicting the one-year ahead operating cash flow, whereas the cash flow-based models has a significantly lower value of 0.039. Takhtaei & Karim (2013) and Ball & Nikolaev (2022) drew the opposite conclusion based on their regression models. Ball & Nikolaev (2022) state that the predictive ability of earnings rises when calculated on an accrual basis, something that is confirmed by El-Sayed Ebaid (2011) and Jemaa et al. (2015). They showed that disaggregating earnings into their major accrual components significantly enhances the predictive performance of earnings on future operating cash flows, with an increase in adjusted R^2 of 8% and 11.3%, respectively. Accruals include income that a firm has earned but has not yet been paid for, e.g., change in inventory.

It should be noted that all of these studies were performed using data of stock-listed, for-profit organizations, and the papers included in the review were predominantly based in the USA. We should therefore be cautious in generalising these results to our situation of a non-profit organisation in the Netherlands. However, based on the above definition of accruals, the monetary value of activities in care paths can be considered accruals, as the expenses for providing care have been incurred, but the hospital has yet to receive payment. A change in inventory translates to our problem context as a change in production capacity of the hospital. Therefore, the production capacity of the hospital might serve as a predictor of future operating cash flows in our study. We discuss different possible prediction models in more detail in Section 3.3.

3.3 Prediction models

As mentioned in Section 3.2, most of the studies included in the Mulenga & Bhatia (2017) review applied regression analyses for assessing the predictors of future cash flow. This is not surprising, as regression models are a convenient, well-known, and long-established statistical methods in forecasting. However, there are more possible statistical methods to consider for developing a forecasting model. Over the last few years, machine learning (ML) methods have been increasingly used for developing prediction models, either by themselves or combined with traditional statistical methods.

Makridakis et al. (2018) evaluated the performance of ML methods as alternatives to statistical ones for time series forecasting in terms of computational requirements, assessed by time needed to train a given model and use it for extrapolation, and by accuracy, measured by forecasting error. Eighteen ML techniques were selected for analysis, including Bayesian neural networks and CART regression trees. After comparison of these eighteen with eight traditional statistical ones, they found that the six most accurate forecasting techniques were statistical, demonstrating their superiority to ML techniques for all investigated forecasting horizons. Additionally, the latter were found to have far higher computational requirements than statistical approaches.

Machine learning methods have also been applied to financial forecasting research. For example, Fatemeh (2020) used artificial neural network methods in addition to linear regression analyses to predict future cash flows. Results showed that the accrual regression model performed better at predicting future cash flow than the other examined models. Tangsucheeva & Prabhu (2014) developed both a Markov chain- and Bayesian model to create a stochastic financial analytics model for cash flow forecasting. According to their results, their stochastic model had a noticeably higher forecast accuracy than the other widely used methods, such as regression models, and were robust to supply chain dynamics.

Combinations of ML and traditional statistic prediction models are broadly applied in other fields of forecasting research as well. Park et al. (2018) analysed the usage and potential of mixed-method approaches in hospitality management. They found that combining both techniques has shown promise for yielding useful insights. In a healthcare setting, Wu & Zhou (2021) assessed the performance of traditional statistical models and machine learning (ML) models by comparing their ability to predict hospital death for patients. Univariate and multivariate logistic regression analyses were used to identify risk factors for hospital death, and Lasso and Boruta methods were used for variable selection, followed by SVM, XGBoost, and ELM as classification algorithms. Results showed that traditional statistics were particularly useful in assessing the relationship between the independent and dependent variables, but the ML model seemed to have a greater clinical utility. Therefore, the authors concluded that both methods are complementary.

Based on the above, no clear consensus regarding the best method for forecasting research exists in scientific literature. This is confirmed by Zaniletti et al. (2023), who assessed the strengths and weaknesses of both traditional and ML models, with a focus on orthopaedic outcome measures. They recommend using traditional statistical models in case of small sample sizes and a limited number of interactions between predictors. If there is no need to interpret the model in detail and overall prediction is the goal rather than isolating the effect of some predictors in the model, ML models are more favourable. Therefore, the most suitable method differs between studies and depends on the research goal, problem context, study design, and data and time constraints. Which method we will use is discussed in Section 4 in more detail.

Section 4

Methodology

In this section, we describe the methodology of this study. First, we define the primary purpose of the prediction model, followed by a solution approach. In Section 4.3, we motivate the study design and data collection. Finally, we describe the statistical analyses to be performed in Section 4.4.

4.1 Aim definition

This study aims at forecasting OCON's operating cash inflows based on capacity indicators. Developing a prediction model has a dual purpose:

- Gaining insight into the factors constituting the cash inflows from operations, and their magnitude and timing.
- Enabling the possibility to adjust operational processes accordingly if considered desirable.

4.2 Solution approach

In this section, we relate the literature from Section 3 to our problem context and research goal. Based on this, we propose a solution approach.

4.2.1 Process prediction

As mentioned before, to the best of our knowledge, only one paper has been published regarding cash flow forecasting in hospitals in the Netherlands: the study by van der Spoel et al. (2013). Therefore, their methods could serve as a guideline for our study. They developed models to predict the activities in care paths and the duration of them, thereby forecasting the billable care product and corresponding monetary value. This means that every DTC care product requires a separate prediction model, which is why they have chosen to focus on three specific diagnoses and corresponding DTC care products.

Developing forecasting models for care products separately instead of one overall prediction model comprising all care products results in more accurate predictions and smaller standard deviations. Applying this method to our situation implies the development of a multitude of forecasting models to predict the overall operating cash flow. Per month, approximately 140 different care products corresponding to over 8,000 patients are opened at OCON. The three most prevalent ones, representing around 1,200 patients, constitute 40% of total operating cash flows, whereas the 5,000 patients representing the top 15 care products correspond to 70%. So, for a proper prediction of the total operating cash flows, at least 15 prediction models should be constructed.

Most patients receive conservative treatments before surgical options are considered. Therefore, to predict the process of patients with for example knee osteoarthritis, we should assess which proportion of patients receives operative treatment within the duration of the care path, and which part is treated conservatively throughout the entire care path. These data are currently not available at OCON, which means that manual analysis of care paths per patient is necessary to acquire this information. In our situation, this means analysing the electronic patient records of at minimum multiple months of 5,000 patients each. Besides potential privacy issues, this is considered not feasible due to time constraints.

Still, it is possible to develop process prediction models for only the three most common care products. However, OCON's financial manager has indicated that it is preferred to have a potentially less accurate model for predicting the overall operating cash flows rather than have more accurate models that predict only 40% of all cash flows. Without an estimate of the remaining 60%, such a model is not suitable in practice. Therefore, process prediction models are not regarded a suitable methodology to answer this study's central research question.

4.2.2 Capacity as a predictor of operating cash flows

As follows from the literature in Section 3.2, future operating cash flows can be predicted based on current cash flows or earnings. It should however be noted that the studies all investigated the ability to predict the cash flow of 1 year in the future, whereas our study aims at predicting the short-term operating cash flow, being several months rather than a year. The reason for this is to enable making business adjustments accordingly, which is regarded most effective in the short-term.

Moreover, OCON has gained financial independence in 2019, meaning there are limited previous cash flow statements to validate the performance of these to predict future cash flows. Because accruals have proven to be predictors of future cash flows, and because capacity serves as a precondition for operating cash flows, production capacity, which underlies all operations, might be suitable as a predictor of future cash flows. Also, capacity is the only possible vessel for steering future cash flows from operations, as OCON for example does not refuse patients with specific diagnoses. To understand the potential predictive abilities of capacity indicators, we provide an explanation of the capacity management at OCON in Section 4.2.2.1. In Sections 4.2.2.2 and 4.2.2.3 we describe the two approaches we will use in our study to forecast operating cash flows.

4.2.2.1 Capacity management

Capacity can roughly be classified in 2 categories: the outpatient clinic capacity that is used to provide conservative care to patients, and the operating room (OR) capacity for performing surgeries. OCON has an outpatient clinic for orthopaedic and sports medical care. All orthopaedic care is included in the basic package of healthcare, which means it is reimbursed by the insurer. This is not the case for the sports medical care, as they provide both insured and non-insured care. An example of insured sports medical care is a patient with exercise-induced complaints. Next to physicians, physician assistants, physical therapists, nursing specialists, and residents consult patients at the outpatient clinic.

If a patient needs a surgery, the anaesthesiology department assesses a patient's pre-operative health state and decides whether they are approved for surgery accordingly. After approval, the OR-date is scheduled by the OR-schedulers. OCON uses 5 operating rooms on Mondays and Thursdays, and 4 operating rooms on the remaining days of the week. During the holiday-related months, the number of used operating rooms and the number of appointments at the outpatient clinic is reduced. After surgery, some patients need to stay one or multiple nights at the clinic to recover. We do not further elaborate on the clinic's capacity here, because we do not regard it relevant for answering our research question as it does not pose a limitation to OR-capacity.

Management and schedulers consider efficient scheduling and high utilization rates important as this increases the number of treated patients, improves patient flow, and reduces idle time. Although both the outpatient clinic and ORs are considered essential, efficient scheduling of the latter has more priority. This has several reasons, the first being that ORs entail higher financial consequences. When used, they generate high cash flows, but they also have high costs, even when unused. Also, an unfilled surgical session has a longer duration than an appointment at the outpatient clinic. That is why outpatient clinic scheduling is more flexible than OR-scheduling, enabling easier short-term scheduling, and is less affected by an unfilled spot in the schedule. Overall, OR-scheduling is more complex.

OR-schedulers perform well at filling the available capacity, ensuring utilization rates remain constantly high throughout the year, even though capacity fluctuates. As the outpatient clinic capacity can be considered a capstone, it makes sense that its utilization rates have more fluctuations. This does not mean that it necessarily is lower, but that it has higher variance than the OR-rates.

4.2.2.2 Forecasting Work-in-Progress

The assumption that production capacity might serve as a predictor of future cash flows is supported by the experience of the financial manager, who has observed that the Work-in-Progress (WIP) follows a similar trend as the production capacity does throughout the year. This is also shown in Figures 3 and 4 (axes are confidential). The data for these figures are extracted from the WIP grouper, a system that monthly extracts the data from OCON's digital environment. WIP represents the monetary value of all care paths that have not been closed yet. Once a care activity is performed, the DTC is immediately included in the WIP. This means there is only small delay in time between the provision of care and the registration of its financial value in WIP. Values are extracted by running the grouper on the opened care paths as if they were closed. Remember, registration of additional healthcare activities or alteration of the typical diagnosis can happen until closure and might result in another care product being returned by the grouper. Therefore, WIP values are inherently uncertain.

Figure 3 shows that the average used OR-hours decreases during the summer holiday-related months, a pattern that can also be seen in the monetary values of WIP. To gain clearer insight into WIP values of the years, we also included Figure 4 that conveys similar information as Figure 3 but excludes the capacity data and includes the WIP data of the year 2020.

To clarify, we provide an example. Imagine patient X. In January, the WIP grouper categorizes his care path as 'conservative treatment knee osteoarthritis'. The grouper only considers activities that have been performed, and not the ones that are scheduled for the future, such as a surgery for patient X. In February, surgery has taken place and patient X's care path is categorized as 'total knee replacement surgery'. Therefore, the WIP grouper has extracted two different DTCs and corresponding financial values in the two subsequent months, and the addition of the surgery has drastically changed the expected cash flow from treating this patient. We do not know which proportion of WIP evolves to different care paths and to which ones they transform. However, as the WIP Grouper dataset monthly comprises care products of over 8,000 patients, we expect this effect to be mitigated by the large sample size. This is supported by the fact that the three most prevalent DTCs represent a relatively constant percentage of total WIP over the months, suggesting constant patient flow. For interpretational purposes, it is important to keep in mind that these data are a snapshot in time.

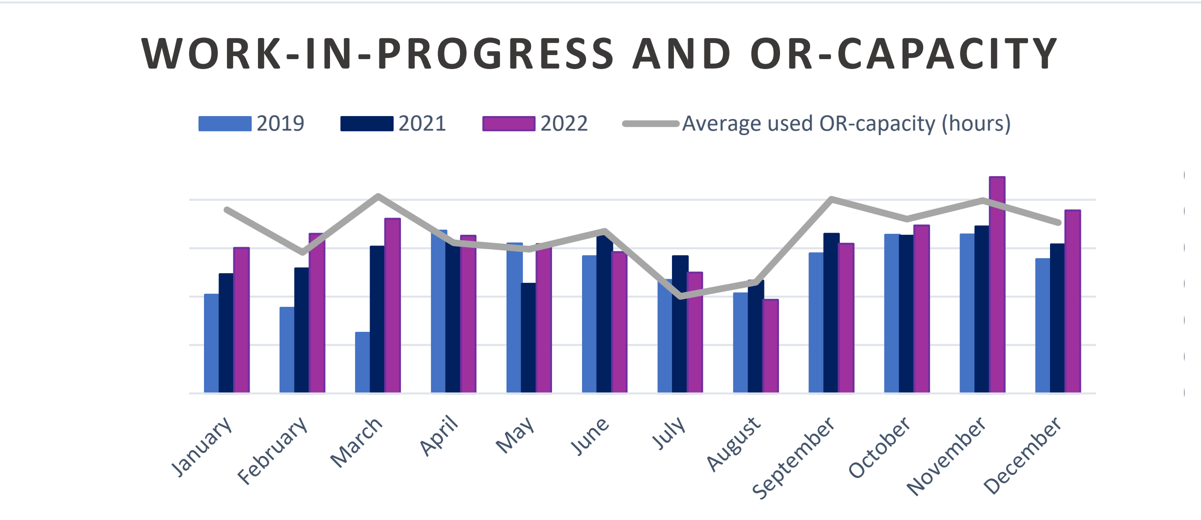

Figure 3. Visualisation of the monetary value of the Work-in-Progress over the year for the years 2019, 2021, and 2022 on the left axis. The right axis shows the used OR-capacity in hours as an average of the same years.

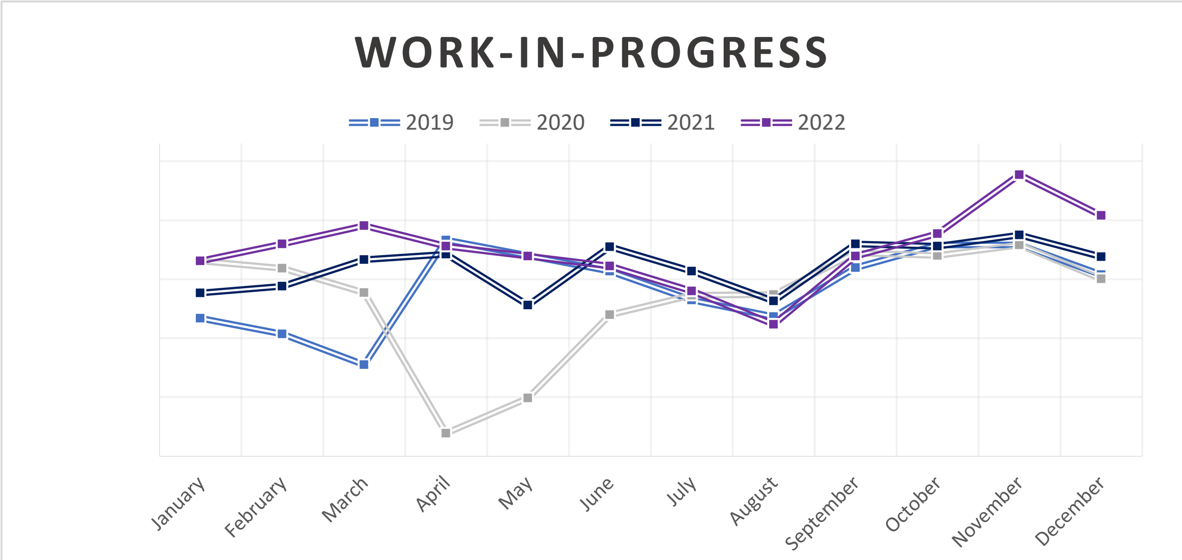

Figure 4. Visualisation of the monetary value of the Work-in-Progress over the year for the years 2019-2022.

Based on the above, WIP can be regarded a precursor of operating cash flows. We expect the WIP to have a more direct correlation with capacity indicators than revenue, as every registered healthcare activity directly is included in the next WIP Grouper data extraction. Figure 4 shows that, for the years 2019, 2021, and 2022, the WIP value decreased during the summer months, followed by an increase during the autumn months and a slight decrease in December. This makes sense, because more present staff means more patients can be treated (assuming constant utilization rates), resulting in more registered health care activities and, subsequently, a higher WIP value. This can also be seen in Figure 3.

The lower WIP values in the months April and May 2020 are explained by the sudden decrease in capacity due the COVID-19 pandemic. Thereby, this decrease supports the assumption of correlation between capacity and WIP. The diversion in the first months of 2019 is considered the result of the transition to financial independence.

In addition to the monetary value of the WIP, the WIP grouper also includes the value of the closed but not yet invoiced care paths, referred to as To-be-Invoiced (TBI). However, these data seem to be highly dependent on the day at which the care path closes, and on the days the grouper extracts, and invoicing takes place. Therefore, these data do not seem to be correlated with production capacity. This can also be seen in Figure 5 (axis is confidential). A peak in the graph can be seen from March till May 2021, because it was not possible to invoice during that period which resulted in an abnormally high TBI values. Although these values are technically correct, they result in a distorted view because they were caused by an unusual situation.

Based on the above, this study aims at forecasting the operating cash flow by using capacity indicators as predictors of WIP. Operating expenses, such as salaries and rent, are not taken into account. This is because OCON already has a good insight into these expenses as they remain rather constant over the year and do not experience an uncertainty in timing and quantity, as is the case with the cash resulting from the provision of care. Forecasting WIP is the first of two approaches we will use to predict operating cash flows at OCON. In Section 4.2.2.3 we describe the second one.

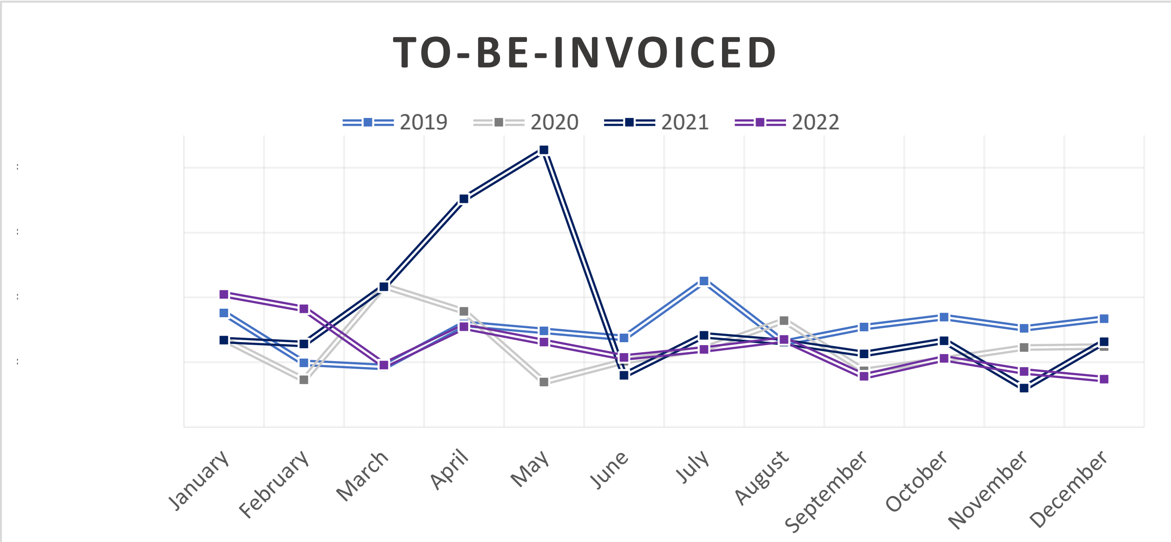

Figure 5. Visualisation of the monetary value of the To-be-Invoiced over the year for the years 2019-2022.

4.2.2.3 Forecasting revenue

When the duration of care path is reached, it closes and WIP turns into TBI, which turns into turnover after invoicing. Therefore, WIP could be considered a precursor of revenue. So, if WIP can be forecasted based on capacity indicators, the same goes for forecasting revenue. We consider this useful because revenue represents the amount of money to-be received from operations. That is why we also aim at forecasting revenue in this study. We should note that cash is considered revenue immediately after invoicing. Therefore, the received amount of cash could be smaller than revenue if debtors fail to pay. This is called default risk.

As described in Section 3, both traditional statistics and machine learning models can be used to develop forecasting models. Which method we should apply in this research is determined based on the specific purpose of our study, which is further elaborated upon in Section 4.3.

4.3 Study design

In this section, we elaborate on the methods used in the study to develop the models for predicting both WIP and revenue in Section 4.3.1 and 4.3.2, respectively. This study has been approved by the ethical review committee at the University of Twente.

4.3.1 Forecasting WIP

4.3.1.1 Data collection and variable selection

For forecasting WIP, we derived data from two different sources. The financial data were pre-existent as the Work-in-Progress Grouper has extracted the data from January 2019 till May 2023. Each data extraction contains the DTC codes corresponding to specific care products and, per DTC code, the number of currently opened care paths per healthcare insurer and their monetary value. Also, an indication whether the path is still open or closed and to be invoiced (WIP vs TBI) is included. For interpretation, it is important to remember that these data are a snapshot in time. The WIP Grouper dataset monthly comprises DTC care products of over 8,000 patients, corresponding to around 140 different care products.

We use Work-in-Progress as the ratio scaled outcome indicator. Although it was not feasible for us to develop separate forecasting models for the different care products as done by van der Spoel et al.

(2013), we divide WIP into conservative and operative care products to decrease the standard deviations of the forecasts. Because of the lack of expected correlation between capacity and TBI, we choose not to include the latter as an outcome indicator.

We extracted production capacity data regarding both outpatient and operating room from a management information system called Business Objects, which directly extracts its data from OCON's digital environment. From these data, we selected six outpatient clinic variables relating to the conservative WIP, five OR variables relating to operative WIP, and 6 general variables to potentially include in our prediction model. These are shown in Table 1.

The conservative WIP predictors are defined as the total number and total duration of scheduled appointments at the outpatient clinic. For example, an orthopaedic surgeon could have fourteen 20-minute appointments and sixteen 10-minute appointments scheduled on a day, resulting in 30 appointments with a total duration of 440 minutes. These variables include combined data on the outpatient clinic appointments of all medical staff over a period of one month and focus only on appointments associated with insured care.

We focus the variables on the time periods one and two months ago and the current month, because we expect these to be correlated to conservative WIP based on care path duration. This is schematically shown in Figure 6a. Remember, conservative care paths have a duration of 90 or 120 days, and care paths are immediately included in WIP after registration of care activities. Therefore, we include variables relating to capacity of maximum two months ago (t-2), which is approximately 120 days before t+1, which means that these care paths could still be included in WIP at t+1. For example, a follow-up care path that was opened on the 3rd of April (t-2) closes 120 days later, on the 1st of August, and therefore is still included in WIP of July (t+1). Please note that these variables relate to scheduled appointments that are scheduled approximately two weeks before they take place. So, we cannot include the scheduled appointments next month as a predictor because most of the appointments have not been scheduled yet. We do not distinguish between appointments that took place or that were last-minute cancelled.

Panel A: Conservative WIP predictors
▪ Number of appointments outpatient clinic current month (NAP_t)
▪ Duration of appointments outpatient clinic current month (DAP_t)
▪ Number of appointments outpatient clinic previous month (NAP_{t-1})
▪ Duration of appointments outpatient clinic previous month (DAP_{t-1})
▪ Number of appointments outpatient clinic two months ago (NAP_{t-2})
▪ Duration of appointments outpatient clinic two months ago (DAP_{t-2})
Panel B: Operative WIP predictors
▪ Available OR-hours next month (AOR_{t+1})
▪ Scheduled OR-hours next month (SOR_{t+1})
▪ Available OR-hours current month (AOR_t)
▪ Scheduled OR-hours current month (SOR_t)
▪ Realised OR-hours current month (ROR_t)
Panel C: General predictors
▪ COVID-19 next month ($COVID_{t+1}$)
▪ COVID-19 current month ($COVID_t$)
▪ COVID-19 previous month ($COVID_{t-1}$)
▪ Reduction next month (RED_{t+1})
▪ Reduction current month (RED_t)
▪ Reduction previous month (RED_{t-1})

Table 1. Variables for predicting Work-in-Progress.

For the OR-variables we include separate variables for the available and scheduled OR-hours. As Figure 6b shows, we use variables related to the OR capacity of the current and next month to forecast operative WIP at t+1. This difference in timing of the variables compared with the outpatient clinic ones is explained by the fact that operative care paths have a shorter duration of maximum 42 days after discharge from the hospital.

Scheduled OR-hours in the next month might seem an odd predictor, as this future value could be uncertain if schedulers have not filled the OR-schedule yet. However, OR-schedules are filled for approximately 80% 12 to 6 weeks in advance. The remainder is reserved for urgency patients and is filled shortly before surgery takes place. Therefore, if scheduled OR-hours proves to be a good predictor of operative WIP, we might need to take measures to create a realistic value of future scheduled OR-hours. Also, realised OR-hours can only be a predictor if assessed at the end of the month. General variables are ones we included in both forecasting models. As can be seen in Figure 4, WIP was greatly decreased during the first wave of COVID-19 when there were many measures to reduce the spread of the coronavirus. To correct for this effect, we included a dummy that indicates if COVID was present or not in the next, current, and previous month. In case of a lockdown in the Netherlands, we consider COVID-19 present, otherwise not. Since we do expect another COVID pandemic, our objective of including this dummy is to improve the accuracy of the model rather than using the dummy for future forecasts (no is the reference level). That is also why it is possible to include the variable "COVID-19 next month".

As mentioned in Section 4.2.2.1, the capacity is reduced during holiday-periods. To account for these changes, we include a dummy variable with three levels. The months February, April, May, October, and December include a short holiday and a corresponding capacity reduction, which is why we have coded these months half reduction. During the summer holiday, in the months July and August, a bigger reduction takes place, which we coded full reduction. The remaining months have no considerable reductions. Similar as with the COVID-19 dummies, we included the reduction dummy for the next, current, and previous month. We do not include the "COVID-19 previous month" and "reduction previous month" in the model for forecasting operative WIP, because of the maximum operative care path duration of 42 days.

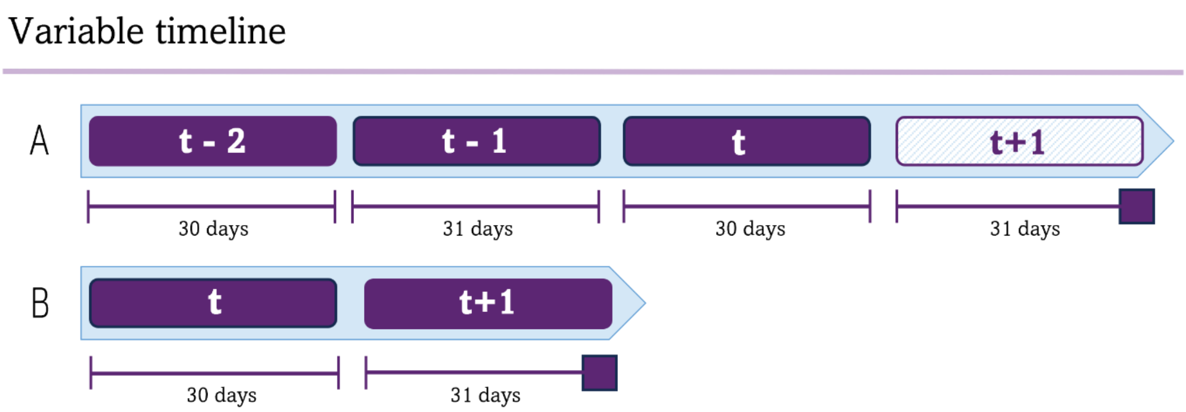

Figure 6. Timeline of the variables used for predicting conservative (A) and operative (B) WIP. The square represents the to-be forecasted WIP.

4.3.1.2 Statistical analyses

In view of the purpose of our study, we have chosen to apply traditional statistical models for forecasting WIP rather than using ML techniques. Arguments for this decision include data considerations, such as the relatively small sample size of our data, and the increased interpretability of statistical methods compared to ML, the lowered computational requirements, and the possibility to isolate the effects of some predictors to allow for production capacity adjustments.

Prior to model construction and analysis, we prepare the dataset by identifying and correcting for missing values, registration errors, and influential observations, and verify the assumptions of linear regression analysis. We use the following forecasting models:

$$\begin{aligned} WIP_{c:\ t+1} = {} & \beta_0 + \beta_1\, OUTPATIENT\ PREDICTOR_t + \beta_2\, OUTPATIENT\ PREDICTOR_{t-1} \\ & + \beta_3\, OUTPATIENT\ PREDICTOR_{t-2} + \beta_4\, COVID_{t+1} + \beta_5\, COVID_t \\ & + \beta_6\, COVID_{t-1} + \beta_7\, REDUCTION_{t+1} + \beta_8\, REDUCTION_t \\ & + \beta_9 REDUCTION_{\ t-1} + \varepsilon_{c:\, t+1} \end{aligned} \tag{1}$$

$$\begin{aligned} WIP_{o:\ t+1} = {} & \beta_0 + \beta_1\, OR\ PREDICTOR_{t+1} + \beta_2\, OR\ PREDICTOR_t + \beta_3\, COVID_{t+1} \\ & + \beta_4\, COVID_t + \beta_7\, REDUCTION_{t+1} + \beta_8\, REDUCTION_t + \varepsilon_{o:\, t+1} \end{aligned} \tag{2}$$

where:

$WIP_{c:\ t+1}$ = Conservative Work-in-Progress next month

$WIP_{o:\ t+1}$ = Operative Work-in-Progress next month

OUTPATIENT PREDICTOR = One of the variables related to the outpatient clinic (Table 1, Panel A)

OR PREDICTOR = One of the variables related to the operating room (Table 1, Panel B)

COVID = Dummy variable for COVID-19 (Table 1, Panel C)

REDUCTION = Dummy variable for capacity reductions (Table 1, Panel C)

$\varepsilon_{\ t+1}$ = Error term of the regression

To clarify the above models, we provide an example. All WIP data were extracted on the last day of the month. So, to predict the conservative WIP of for instance the 30th of June, we include the outpatient clinic predictors number and duration of appointments in May, April, and March and the COVID-19 and reduction dummies for the June, May, and April. To forecast operative WIP of the 30th of June, we use the OR predictors available and scheduled hours June, and the available, scheduled and realised hours of May. We also include the COVID-19 and reduction variables for the months June and May. Remember, operative care paths have a maximum duration of 42 days after discharge from the hospital, which is why we do not include OR-variables of further in the past.

Although we explore the predictive abilities of all variables from Table 1, it is unlikely they will all be included in the final model. We use the backward stepwise regression method to identify the combination of variables resulting in a reduced linear regression model that best predicts the WIP. Step by step, the least significant variables are excluded by the model, until a model with only significant independent variables ($p < 0.05$) remains. The combination of significant variables resulting in the model with the highest adjusted R^2 is considered the best forecasting model. Besides, we do not consider including all variables a desirable situation, as this would result in multicollinearity between some variables. For example, the number and duration of appointments convey similar information.

We initially include them both with the goal of identifying the best predictor of WIP. We aim at including a maximum of 5 independent variables to increase interpretability and because of the relatively small sample size.

All statistical analyses are performed using the R software (version 4.2.3). P-values of 0.05 or lower are considered statistically significant. After model construction, we assess if the remaining four assumptions of regression analysis, being normality, independence, uncorrelatedness, and constant variance of the error terms are met. We evaluate variance inflation factors (VIF) to estimate how much the variance of the regression coefficients is inflated due to multicollinearity.

To assess external validity, we randomly split the dataset into a training and test set with an 80:20 ratio. We select variables and train the model based on the training dataset. We evaluate model fit considering the root mean squared error (RMSE) as absolute measure, and adjusted R-squared (R^2) as relative measure. We use adjusted R^2 instead of R^2 because it incorporates the model's degrees of freedom, thereby only increasing its value when adding predictors to the model increases model fit more than it decreases the degrees of freedom. We consider adjusted R^2 values from 0.50 acceptable, where higher values are preferred. The F-statistic tells us whether the proposed relationship between WIP and predictor variables is statistically reliable. We assess forecast accuracy by plotting the predictions against the observed values, and by calculating the forecast errors as a percentage of the observed values. We perceive accuracy percentages of over 70% acceptable and over 90% as good.

Then, we use the developed model to predict the WIP for the test data. We assess model performance in the same way we evaluated the model on de training data. As the test dataset contains 20 percent of our total dataset, this results in a rather small test sample. Therefore, potential outliers and chance have considerable impact on the results, and there is limited statistical power to accurately test the quality of the model. Therefore, conclusions regarding generalizability should be drawn with caution.

4.3.2 Forecasting revenue

4.3.2.1 Data collection and variable selection

We use revenue as the ratio scaled outcome indicator and include only revenue from insured healthcare activities. The data were pre-existent and provided to us by OCON's financial manager. In 2019, revenue was registered on a yearly basis. From January 2020 till September 2021 revenue was quarterly reported, and from October 2021 till May 2023 there are monthly revenue reports.

Because we want to make monthly revenue forecasts, we need to transform the quarterly data to monthly data. We do this by dividing the quarterly data by three, resulting in the same revenue for the three months in the same quartile. Under the assumption of a stationary distribution, we consider the standard deviation of revenue to be constant for all observations. Therefore, we can model monthly revenue values from the quartile data following four steps for each month separately. First, we generate a random number from a standard normal distribution. We choose a normal distribution because we assume that revenue is normally distributed. Secondly, we multiply that random number with the standard deviation of the monthly observations. Then, we add this value to the average of the quartile data. Finally, we perform 100 iterations of this process and then take the average of all iterations, leaving us with an estimated value of revenue that matches the standard deviation of the overall sample. In Section 5.2.1, we provide an example of one of these calculations.

Similar to the data collection for predicting WIP, we extracted the data regarding capacity measures from Business Objects. From these data, we selected six outpatient clinic variables, six OR variables, and ten general variables to potentially include in our prediction model. These are shown in Table 2.

As mentioned in Section 4.2.2.3, we only receive revenue after care paths are closed and invoiced. Therefore, it takes a longer time for capacity to be transformed into revenue than into WIP. After care

Panel A: Predictors of revenue related to the outpatient clinic
▪ Number of appointments outpatient clinic two months ago (NAP_{t-2})
▪ Duration of appointments outpatient clinic two months ago (DAP_{t-2})
▪ Number of appointments outpatient clinic three months ago (NAP_{t-3})
▪ Duration of appointments outpatient clinic three months ago (DAP_{t-3})
▪ Number of appointments outpatient clinic four months ago (NAP_{t-4})
▪ Duration of appointments outpatient clinic four months ago (NAP_{t-4})
Panel B: Predictors of revenue related to the operating room
▪ Available OR-hours current month (AOR_t)
▪ Scheduled OR-hours current month (SOR_t)
▪ Realised OR-hours current month (ROR_t)
▪ Available OR-hours previous month (AOR_{t-1})
▪ Scheduled OR-hours previous month (SOR_{t-1})
▪ Realised OR-hours previous month (ROR_{t-1})
Panel C: General predictors
▪ COVID-19 current month ($COVID_t$)
▪ COVID-19 previous month ($COVID_{t-1}$)
▪ COVID-19 two months ago ($COVID_{t-2}$)
▪ COVID-19 three months ago ($COVID_{t-3}$)
▪ COVID-19 four months ago ($COVID_{t-4}$)
▪ Reduction current month (RED_t)
▪ Reduction previous month (RED_{t-1})
▪ Reduction two months ago (RED_{t-2})
▪ Reduction three months ago (RED_{t-3})
▪ Reduction four months ago (RED_{t-4})

Table 2. Variables for predicting revenue.

paths have closed, a grouper run and invoicing generally takes place after 10 to 16 days. However, registration errors can result in faulty, non-claimable care product extractions the grouper. These errors must be corrected, which means it takes at least one additional week before invoicing can take place.

To investigate what variables from which months should be included in our forecasting model, we visualized a timeline in Figure 7. In this figure, we focus on the example of forecasting revenue for till September. It is important to remember that revenue includes all invoices that were sent from the first the last day of the month. Due to uncertainty in the duration of care paths and time between grouper run and invoice, Figure 7 includes four different scenarios.

Figure 7a shows the longest possible conservative care path included in the revenue in September. In this example, the follow-up care path is opened at the 18th of April and closes 120 days later, on the 16th of August. The time between care path closing and invoicing depends on the day of the week that the care path has closed but has a maximum duration of sixteen days. Therefore, this invoice is sent on the 1st of September and thus is included in September's revenue. Similarly, Figure 7b shows the shortest possible conservative care path. In this situation, an initial care path is opened on the 22nd of June and closed 90 days later, on the 20th of September. Invoicing takes place after the minimum duration of ten days after care path closure, which results in an invoice sent on the 30th of September. Based on these results, conservative care paths opened between the 18th of April and the 22nd of June can be included in September's revenue, which is why we include the outpatient clinic variables of April (t-4), May (t-3), and June (t-2) as predictors.

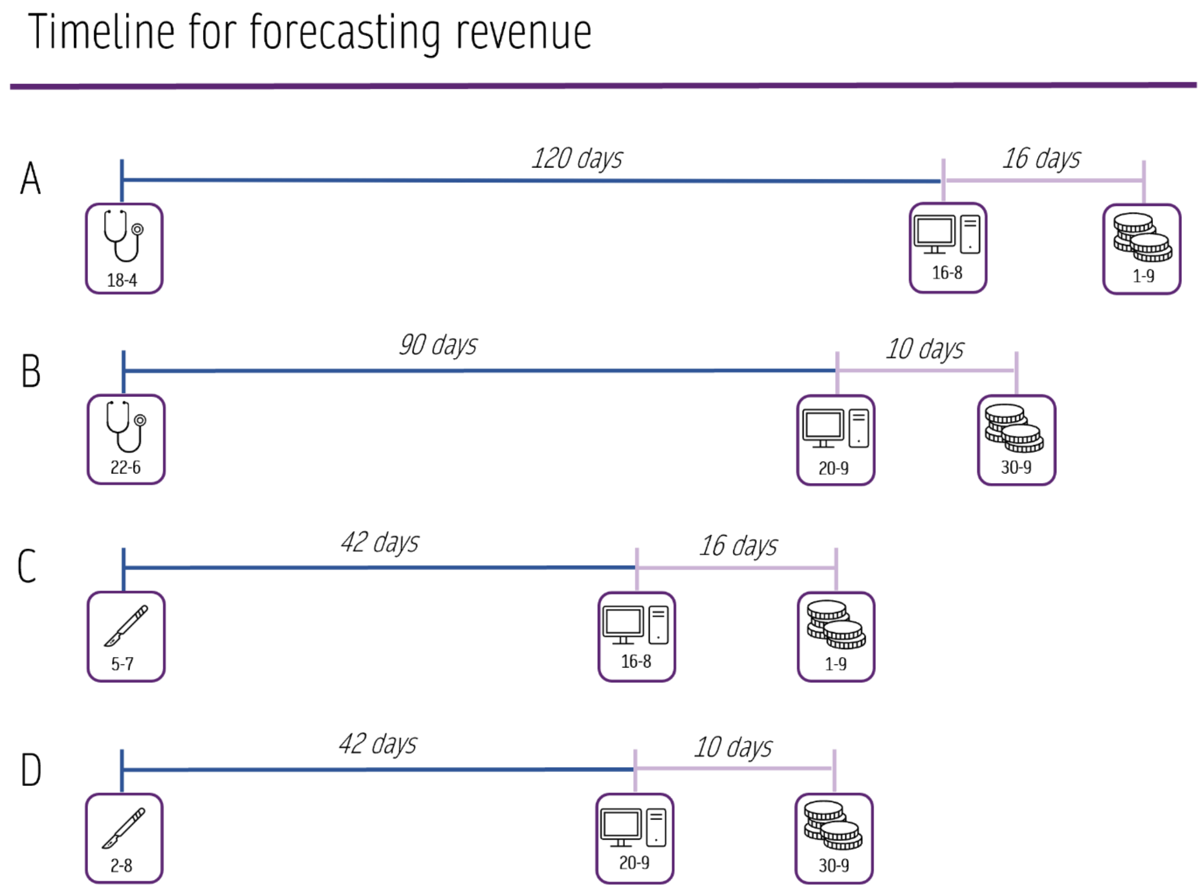

Figure 7. Timeline for forecasting revenue using four possible scenarios. This figure focuses on the example of forecasting revenue of September. A stethoscope icon represents the opening of a care path, and a computer icon represents closing of the care path and a run by the grouper. The coins icons indicate sending of an invoice.

Predictors related to operative care paths are identified in the same way, but the difference is the care path duration of 42 days after discharge from the hospital. This can be seen in Figure 7c and 7d. Surgeries performed between the 5th of July and the 2nd of August are expected to contribute to September's revenue, which is why we include the OR-variables of July (t-1) and August (t). This reasoning can be generalized to other months, even though Figure 7 and the above-described example focus on forecasting revenue of September. The general timeline for forecasting revenue is shown in Figure 8.

Although we account for the vast majority of possible care paths, please note that there can be paths not included by the above-described variables. Examples include paths with a longer duration between closing and invoicing of the care path due to registration errors or closure of care paths that close very shortly after surgery because the maximum path duration of 120 days is reached. We also include dummy variables representing COVID-19 and reduction for every period included in the other variables.

Because of the longer timespan between providing care activities and receiving revenue than the time between care activities and inclusion in WIP, we expect a lower performance of the model forecasting revenue than the model forecasting WIP. After all, an increased timespan is associated with increased variability in processes, thereby complexifying forecasting. This could result in forecasts with lower accuracy and larger standard deviations. On the other hand, revenue forecasts are based on the care that was truly rendered, whereas WIP forecasts, specifically conservative WIP, are primarily based on data of previous months.

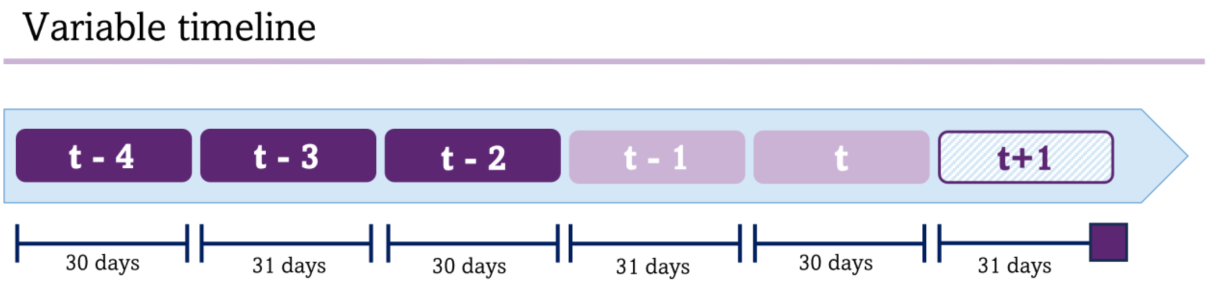

Figure 8. Timeline of the variables used for predicting revenue. The square represents the to-be forecasted revenue. The dark purple rectangles represent outpatient clinic variables, and the lilac ones represent OR-variables.

4.3.2.2 Statistical analyses

We have chosen to apply linear regression models to forecast revenue. Before we construct the models, we will clean the data by identifying and repairing any missing values, influential observations, and registration errors. We will develop forecasting models using both the raw an adjusted data, that comprises the generated monthly revenue values as described above. Unlike WIP, we cannot split revenue into a conservative and operative part. Therefore, we will combine all variables in one forecasting model. We use the following forecasting model (Model 3):

$$\begin{aligned} REV_{t+1} = \ & \beta_0 + \beta_1 \ OUTPATIENT\ PREDICTOR_{t-2} + \beta_2 \ OUTPATIENT\ PREDICTOR_{t-3} \\ & + \beta_3 \ OUTPATIENT\ PREDICTOR_{t-4} + \beta_4 \ OR\ PREDICTOR_t \\ & + \beta_5 \ OR\ PREDICTOR_{t-1} + \beta_6 \ COVID + \beta_7 \ REDUCTION + \varepsilon_{t+1} \end{aligned} \quad (3)$$

where:

REV_{t+1} = Revenue next month

OUTPATIENT PREDICTOR = One of the variables related to the outpatient clinic (Table 2, Panel A)

OR PREDICTOR = One of the variables related to the operating room (Table 2, Panel B)

COVID = Dummy variable for COVID-19 (Table 2, Panel C)

REDUCTION = Dummy variable for capacity reductions (Table 2, Panel C)

ε_{t+1} = Error term of the regression

To improve clarity, we did not include the COVID-19 and reduction variables for all months separately in the above formula. However, in the final model, it is possible that multiple months of either of these dummy variables are included. Model 3 should be interpreted as follows. To predict the revenue of August, we include the outpatient variables number and duration of appointments in March, April, and May, the available, scheduled, and realised OR-hours of June and July, and the COVID-19 and reduction variables for the months March, April, May, June, and July.

Similar to the WIP forecasting models, we initially include all variables to avoid possibly omitting relevant variables but will not include them all in the final prediction model. Variable selection, regression analysis, and performance assessment methods and cut-off points are equal to the methods described in Section 4.3.2.1. There is one exception, as we do not evaluate external validity of the model due to the small sample size.

Section 5

Results

In this section, we describe the results of the forecasting models. In Section 5.1 and its subsections, we focus on forecasting WIP by preparing the data, inspecting the descriptive statistics, selecting variables, and assessing model performance. In Section 5.2, we do the same for the models predicting revenue.

NOTE: some results are excluded because of confidentiality. The confidential annex is registered in the repository of the University of Twente.

5.1 Forecasting WIP

5.1.1 Data preparation and descriptive statistics

Prior to model construction and analysis, we prepared the dataset, and identified and corrected for missing values, registration errors, and influential observations. To allow splitting of WIP data in the conservative and operative subsets, we transformed the DTC codes from in the WIP grouper data into care product descriptions. We deduced whether care products and their corresponding WIP value relate to conservative or operative treatments accordingly. Some care products relating to the Ear, Nose & Throat (ENT) medical speciality were found in the 2019 and 2020 data. Those have been excluded. We also identified incorrect values in the outpatient clinic capacity data. Blockages in schedules were counted as appointments, resulting in an overestimation of the number and duration of appointments at the outpatient clinics. No value-creating activities had taken place during these blockages, which is why we excluded those. The above errors are the logical consequence of the fact that data are directly extracted from OCON's digital environment, making their quality dependent on the quality of registration by users.

Table 3 provides a summary of descriptive statistics of the variables used in this study. No COVID-19 and no reduction are the reference levels for the dummy variables. The raw data show two missing values in the WIP variables that are caused by the fact that the WIP grouper did not extract data in May 2019 and 2022. We chose to delete those observations instead of estimating them, because of the limited amount of reference data to base the estimates on, specifically because both missing values are from the month May. That is why a potentially faulty estimation would have a higher negative impact on the data quality than the positive effect of including two more observations.

Regression analysis makes six assumptions, two of which should be assessed prior to model construction. We check for the presence of a linear relationship between the dependent and independent variables by using scatter- and added variable plots. Those can be seen in Appendix A and B. Scatter plots show the relationship between the dependent and independent variables, whereas added variable plots also consider the effect of the other independent variables in the model. They estimate the relationship between the response variable and an independent variable, given the effect of the other independent variables in the model. In both types of plots, for conservative and operative WIP, we do not observe non-linear patterns. Some independent variables show a more evident linear relationship than others, but we still include all variables in the model because we do not want to omit potentially relevant variables in advance. Added variable plots showed influential observations for January, February, and March 2019 for both the conservative and operative WIP on multiple variables. These might be explained by errors in Grouper data extractions or as a valid but exceptional observation, both explained by the transition to financial autonomy in the beginning of 2019. We chose to delete those variables from the dataset because of their large, presumably incorrect impact on our results. Deletion of the missing and influential observations resulted in our final dataset, of which the descriptive statistics are shown in Table 3 (excluded because of confidentiality). Final data have slightly higher mean and median conservative and operative WIP values and smaller standard deviations. Data preparation excluded the minimum conservative WIP value, resulting in a higher minimum value in the final data. No considerable changes in the independent variable summary statistics were observed.

	$WIP_{c,t+1}$	NAP_t	DAP_t	NAP_{t-1}	DAP_{t-1}	NAP_{t-2}	DAP_{t-2}
$WIP_{c,t+1}$	1.00						
NAP_t	0.53*** (0.29-0.71)	1.00					
DAP_t	0.53*** (0.29-0.71)	0.95*** (0.91-0.97)	1.00				
NAP_{t-1}	0.33** (0.05-0.56)	0.06 (-0.22-0.34)	0.12 (-0.17-0.39)	1.00			
DAP_{t-1}	0.38*** (0.11-0.60)	0.18 (-0.11-0.44)	0.31** (0.03-0.55)	0.94*** (0.90-0.97)	1.00		
NAP_{t-2}	-0.02 (-0.30-0.27)	0.00 (-0.29-0.28)	0.07 (-0.22-0.35)	0.06 (-0.23-0.34)	0.12 (-0.17-0.39)	1.00	
DAP_{t-2}	0.09 (-0.20-0.37)	0.09 (-0.20-0.37)	0.25* (-0.04-0.50)	0.17 (-0.12-0.43)	0.32** (0.04-0.55)	0.94*** (0.90-0.97)	1.00

	$WIP_{o,t+1}$	AOR_{t+1}	SOR_{t+1}	AOR_t	SOR_t	ROR_t
$WIP_{o,t+1}$	1.00					
AOR_{t+1}	0.50*** (0.25-0.69)	1.00				
SOR_{t+1}	0.81*** (0.69-0.89)	0.66*** (0.46-0.80)	1.00			
AOR_t	0.20 (-0.08-0.46)	0.01 (-0.27-0.29)	-0.03 (-0.31-0.26)	1.00		
SOR_t	0.61*** (0.39-0.76)	0.21 (-0.08-0.46)	0.34** (0.06-0.57)	0.67*** (0.48-0.80)	1.00	
ROR_t	0.56*** (0.33-0.73)	0.20 (-0.09-0.46)	0.31** (0.03-0.55)	0.69*** (0.51-0.81)	0.99*** (0.97-0.99)	1.00

Table 4. Correlation (95%-CI) between the variables for predicting conservative and operative WIP.* Correlation is significant at the 0.1 level, ** at the 0.05 level and *** at the 0.01 level.

We check for the assumption of no perfect multicollinearity by assessing the correlation matrix, which is shown in Table 4. The significance indicates whether the correlation between two variables significantly differs from zero. The confidence intervals (CI) show us the range in which the true correlation lies with 95% confidence. Between the independent variables, significant correlations of over 0.8 with a narrow CI surrounding the high correlation are considered problematic.

The scheduled and realised OR-hours have a significant correlation of 0.99 with a narrow confidence interval, suggesting the high probability of a truly strong correlation between those variables. This is caused by the fact that OR-schedulers accurately estimate surgical duration based on previous realised OR-hours. Similarly, there are high correlations with small CIs between the variables number and duration of appointments at the outpatient clinic in the same month, which was expected as they convey the same information. To not violate the assumption of no perfect collinearity, these combinations of variables cannot be included simultaneously in the forecasting models. Therefore, their forecasting potential is examined in separate models to select the variables with the best predictive ability accordingly.

5.1.2 Forecasting models

In this section, we construct the models for forecast conservative and operative WIP. To explore the predictive power of the variables from Table 1, we investigate all possible subsets of independent variables.

5.1.2.1 Conservative WIP

This section comprises two parts. First, we select the variables and construct the model, followed by an assessment of model performance.

Model 1: $\boldsymbol{WIP_{c:\,t+1} = \beta_0 + \beta_1 NAP_t + \beta_2 NAP_{t-1} + \beta_3 COVID_{t+1} + \beta_4 COVID_t + \beta_5 REDUCTION\ (full)_{t+1} + \beta_6 REDUCTION\ (half)_{t+1} + \varepsilon_{c:\,t+1}}$

	Coefficient (Standard Error)	*T-value*
NAP_t		5.04***
NAP_{t-1}		3.91***
$COVID_{t+1}$ (yes)		-2.80***
$COVID_t$ (yes)		-2.92***
$Reduction_{t+1}$ (full)		-3.73***
$Reduction_{t+1}$ (half)		-1.87*
Constant		2.55**
Adjusted R^2	0.71	
RMSE		
F	16.95	
Sig.	0.00	

Table 5. Model for predicting the conservative Work-in-Progress. Significance is indicated by asterisks, where *** indicates a significance at the 0.01 level, ** at the 0.05 level and * at the 0.1 level.

Variable selection

Splitting the data has resulted in a training dataset comprising forty observations of conservative WIP and the independent variables. As indicated in Section 5.1, we should not include the variables number and duration of appointments of the same month in one model. Because a combination of these variables would result in non-significant variables, these models are automatically excluded as potential best model since we require all independent variables to be significant. Backward variable selection results in Model 1 and includes the predictor variables number of outpatient clinic appointments of the current and previous month, the COVID dummy for the current and next month, and the reduction dummy for next month. The results are presented in Table 5 (coefficients are excluded because of confidentiality).

In Model 1, the estimates are compliant with what could be expected based on logical reasoning. They indicate that WIP increases when the number of appointments at the outpatient clinic rises and decreases in times of capacity reductions or COVID. All predictor variables and the constant are significant, besides the dummy representing months with half reduction (p=0.07). Ideally, we would want to remove the half reduction variable from the model, but it is not possible to include some significant categories of a variable and exclude the others that do not significantly differ from zero. Therefore, removal of the half reduction variable would also result in the deletion of the full reduction one, which significantly and considerably impacts the WIP forecast based on the results of Table 5. Because we consider the added value of the full reduction variable to the model more important than the impact of including a non-significant variable in the model, we chose to include the Reduction variable in its entirety in Model 1.

Model performance

Before we can assess model performance, we should verify whether the assumptions of linear regression are met. We use generalized variance inflation factors (GVIF) instead of straightforward variance inflation factors (VIF), because these correct for the higher VIF values for variables with more than one degree of freedom, such as dummy variables with more than two categories. In Model 1, this is the Reduction variable. The GVIF values are shown in Table 6. All predictors have GVIFs close to one, indicating diminutive correlation between the independent variable that is negligible for interpreting the results of the regression analysis.

Figure 9 shows the residual scatter plot for Model 1 (x-axis excluded because of confidentiality). We observe no pattern, suggesting homoscedastic and uncorrelated error terms. Additionally, the Breusch-Pagan test was not significant (p=0.61), meaning we do not have evidence that heteroscedasticity is present in Model 1. In the Q-Q plot of Figure 10 (y-axis excluded because of confidentiality), almost all points are approximately aligned along the reference line, which is why we assume normal distribution of the error terms. We contribute the one observation that falls outside the bandwidth to chance. The final assumption of regression analysis, independence of the error term, is difficult to investigate based on statistical means and is usually judged based on theoretical reasoning. As the error term includes all variables that have not been included in the model, omitting relevant variables could result in dependence of the error term. Because we have explored the duration and number of appointments at the outpatient clinic for several months in our model to find the best predictors of conservative WIP, we reason that the error term is independent.

	GVIF
NAP_t	1.19
NAP_{t-1}	1.06
$COVID_{t+1}$	1.36
$COVID_t$	1.35
RED_{t+1}	1.11

Table 6. Generalized variance inflation factors for Model 1.

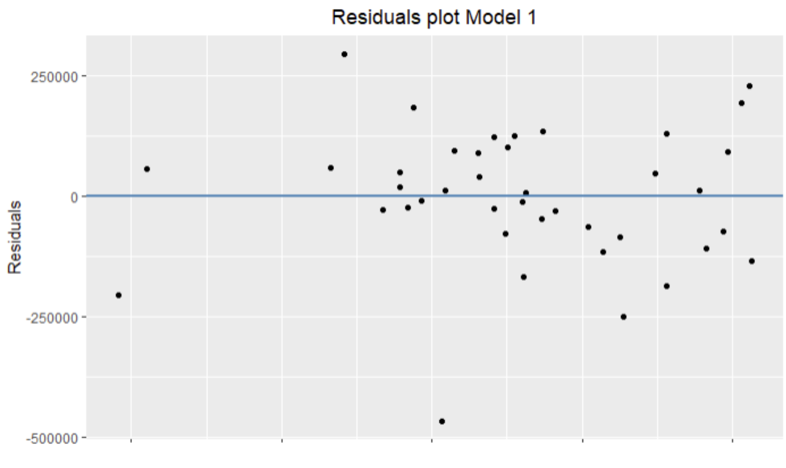

Figure 9. Residual scatter plot for Model 1 on the training dataset.

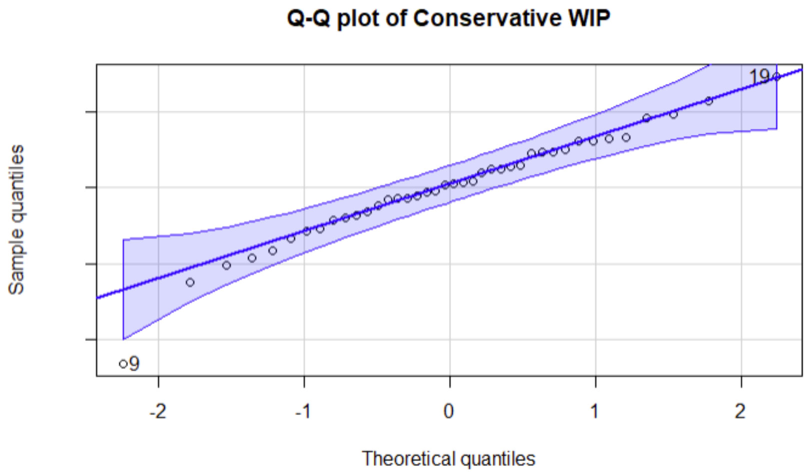

Figure 10. Q-Q plot for Model 1 on the training dataset.

We analyse goodness of fit as a measure of model performance. Adjusted R^2 of Model 1 is 0.71, indicating that 71 percent of the variance in conservative WIP is explained by the predictors in the model. The F-test is significant, indicating that the observed adjusted R^2 is significant. By comparing the observed and predicted values of WIP, we calculated the forecasting error. The mean, median, minimum, and maximum forecast accuracy of Model 1 on the training data are 94.9%, 96.2%, 70.0%, and 99.6%, respectively. This minimum accuracy value corresponds to the prediction of conservative WIP of May 2020. This was during the first COVID wave, which might explain this lowered accuracy, even though we included the COVID variable. Without this observation, the minimum accuracy was 95.5%. This outlier also explains the lower average than median of forecast accuracy.

To assess generalizability, we perform an out-of-sample evaluation on the test dataset. R^2 of the test dataset is 0.49. This indicates that application of Model 1 on the test data results in a lower percentage of explained variance of WIP and a larger standard deviation of the residuals compared with Model 1 on the training data. However, we should be cautious in drawing conclusions based on the results of the test set, due to its small sample size comprising 8 observations.

In Figure 11 (figure excluded because of confidentiality), the predictions of WIP are plotted against the observed values. In case of perfectly accurate predictions, all dots would be aligned on the blue line meaning that the predictions equal the observed values. So, smaller distance between the dots and the blue line indicates more accurate predictions. We cannot evaluate forecasting error based on Figure 11, as there are too few dots to identify a pattern. In Table 7 (excluded because of confidentiality), we calculated forecasting accuracy of all predictions from the dataset. The mean, median, minimum, and maximum forecast accuracy on the test data are 93.6%, 95.3%, 89.5%, and 96.7%, and indicate a slight lowered forecast accuracy for the model on the test data than on the training data. However, true generalizability could be higher or lower because chance highly influences the results in a test dataset of 8 observations.

5.1.2.2 Operative WIP

Like Section 5.1.2.1, in the first part of this section we select the variables and construct the model. Then, we assess model performance.

Variable selection

After splitting the dataset, we have a training and test dataset of forty and eight observations, respectively. Backward variable selection resulted in Model 2, including the variables scheduled OR hours current and next month, and the dummy variable COVID next month. As required, this model does not include both scheduled and realised hours current month. The results of Model 2 are presented in Table 8 (some results excluded because of confidentiality).

Like Model 1, the estimates of Model 2 are intuitive. They indicate that operative WIP increases when the number of scheduled OR-hours rises and decreases in times of COVID. All predictor variables are significant at the 0.01 or 0.05 level, besides the constant (p=0.87). As with all other non-significant variables, the constant could be removed from the model. The absence of a constant suggests that the regression line should pass through the origin. This seems to make sense, because performing no surgeries would result in no operative WIP. However, we cannot draw this conclusion. Although it is an unlikely situation, some surgeries performed at t-1 can still be included in the operative WIP of t+1. This is for example the case when a patient has to remain hospitalized for a longer period of time after surgery, e.g. because of post-operative complications. Remember, the care path closes 42 days after discharge instead of 42 days after surgery.

Model 2: $\boldsymbol{WIP_{o:\,t+1} = \beta_0 + \beta_1\, SOR_{t+1} + \beta_2\, SOR_t + \beta_3 COVID_{t+1} + \varepsilon_{o:\,t+1}}$

	Coefficient (Standard Error)	*T-value*
SOR_{t+1}		8.82***
SOR_t		4.62***
$COVID_{t+1}$ *(yes)*		-2.35**
Constant		0.18
Adjusted R^2	0.82	
RMSE		
F	60.23	
Sig.	0.00	

Table 8. Model for predicting the conservative Work-in-Progress. Significance is indicated by asterisks, where *** indicates a significance at the 0.01 level, ** at the 0.05 level and * at the 0.1 level.

Even if it is uncertain whether the regression line should go through the origin, insignificance could still be a cause to delete the constant from the model. After all, this indicates that the value does not significantly differ from zero. Deleting the constant from Model 2 results in the variable $COVID_{t+1}$ losing its significance and an adjusted R^2 of 99.3%. By removing the constant, the adjusted R-squared has become an unreliable measure of model performance (UCLA: Statistical Consulting Group, 2023).

This is explained as follows. Removing the constant indicates that the expected value of WIP is 0 when all predictors equal 0. As stated above, this assumption might be incorrect. If the constant does not have a true zero, the sum of squares of the model increases relatively more than the sum of squares of residuals. Consequently, (adjusted) R^2 increases. This suggests that Model 2 does not pass through the origin. Because of this, and because we highly value adjusted R^2 as a measure of model performance, we choose to include the constant in Model 2 although it being non-significant.

Model performance

Since we did not include a variable with more than one degree of freedom, we can use VIF values to detect the potential presence of multicollinearity in Model 2. These are shown in Table 9. All predictors have VIFs that are near to one, which indicates a negligible correlation between the independent variables for interpreting the model results. The residual scatter plot for Model 2 is shown in Figure 12 (x-axis excluded because of confidentiality). Since we do not discover a pattern in the dots, we assume homoscedastic and uncorrelated error terms. Additionally, the Breusch-Pagan test returned a non-significant result (p=0.86), leading us to accept the null hypothesis of homoscedasticity. Figure 13 shows the Q-Q plot (y-axis excluded because of confidentiality). As almost all points are scattered along the reference line, we assume that the error terms have a normal distribution. We argue that the error term is independent since we included various OR capacity measures in our model and controlled for the influence of capacity reductions and COVID to determine the best predictors of operative WIP.

	VIF
SOR_{t+1}	1.24
SOR_t	1.16
$COVID_{t+1}$	1.10

Table 9. Variance inflation factors for Model 2.

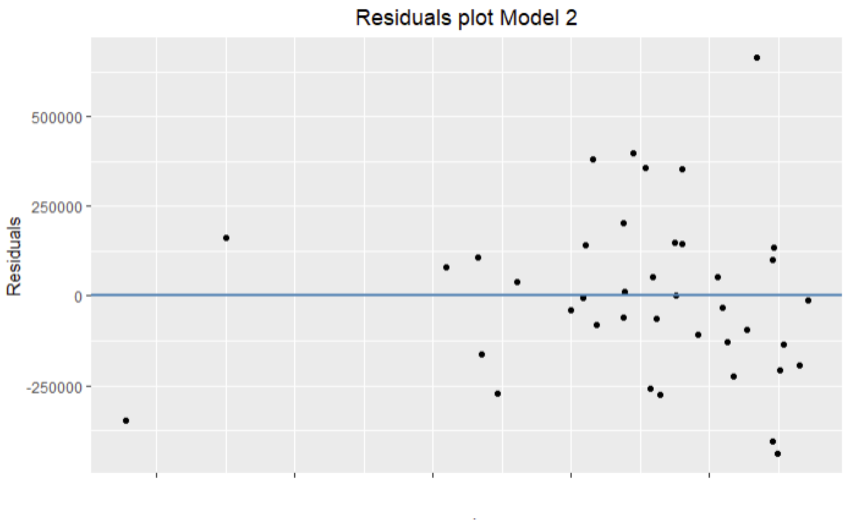

Figure 12. Residual scatter plot for Model 2 on the training dataset.

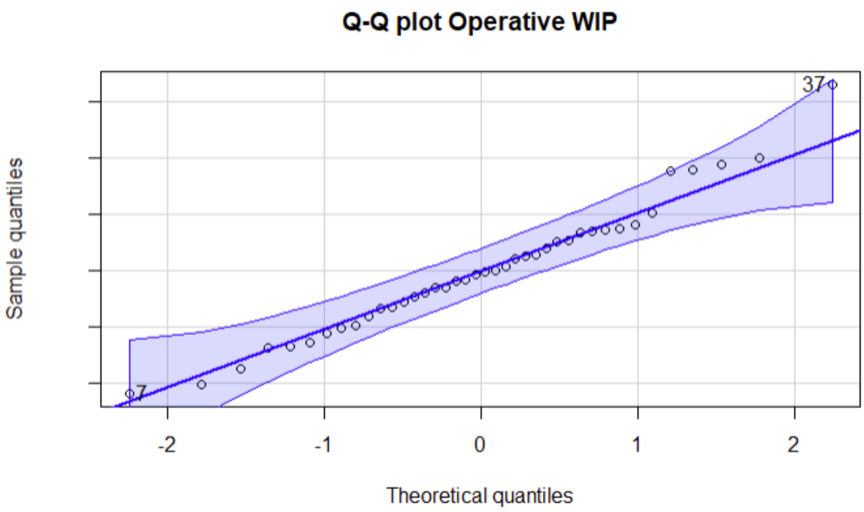

Figure 13. Q-Q plot for Model 2 on the training dataset.

To assess goodness of fit, we evaluate adjusted R^2 and RMSE. Adjusted R^2 is 0.82 and considered to be significant because of the significant F-test. The mean, median, minimum, and maximum forecast accuracy of Model 2 on the training data are 92.4%, 95.2%, 35.6%, and 99.9%, respectively.

This minimum value corresponds to the prediction of WIP of April 2020, which was during the first COVID wave. Remember, the minimum forecast accuracy of Model 1 was also in this period. This makes sense because operative care paths have a smaller lead time than conservative ones. Therefore, the measures for COVID more quickly impact operative WIP, leading to an abnormally low WIP in April 2020. Model 2 was not able to accurately predict this value, despite the inclusion of the COVID variable in the model. When we exclude the prediction of April 2020, the minimum accuracy is 82.7%. The average then rises to 93.9%.

Applying Model 2 to a test dataset provides us with a measure of generalizability. R^2 of the test dataset is 0.82, indicating that the percentage of explained variance of WIP remains unchanged and standard deviation of the residuals increases when Model 2 is used to predict new data. However, we should be cautious in drawing conclusions based on the results of the test set, due to its small sample size comprising 8 observations. In Figure 14 (excluded because of confidentiality), we plotted the observed values of WIP against the predictions. Besides one outlier, the dots appear to be equally close to the line. We attribute this outlier to chance. However, similar as with Model 1, we cannot accurately evaluate forecasting error based on Figure 14, as there are too few dots to identify a pattern.

Table 10 (excluded because of confidentiality) shows the forecasting accuracy of all predictions from the dataset. The mean, median, minimum, and maximum forecast accuracy on the test data are 94.8%, 96.4%, 83.8%, and 99.6%, respectively. These values seem slightly higher than the forecast accuracy on the training data, with less variability. This result is most likely explained by the smaller size of the test data compared to the training set. Also, as chance has a significant impact on the outcomes in a dataset of 8 observations, true generalizability could differ from these values.

5.2 Forecasting revenue

5.2.1 Data preparation and descriptive statistics

Table 11 shows the descriptive statistics for revenue and the variables used for forecasting (excluded because of confidentiality). No COVID-19 and no reduction are the reference levels for the dummy variables. In the first few months of a new year, insurers have not digitally approved the newly negotiated DTC prices yet. Until they have done so, it is not possible to invoice care paths that were opened in 2023. As a result, TBI is relatively high during these first months and revenue relatively low. Usually, this problem is solved in May, resulting in an exceptionally high revenue in that month, and lower revenues in the months before. Although these values are technically correct, they are caused by an uncommon situation, making them influential observations. However, we cannot delete these observations because it is an annual recurring issue and because we need these observations to increase sample size. We did not identify any other influential observations, missing values or registration errors in the raw data.

Originally, de data set comprises 20 observations, because monthly revenue reporting started from October 2021 on. As mentioned in Section 4.3.2, we transform the quarterly data to monthly data to increase sample size. We provide one example of how we did that in Table 13 and the text below (adjusted because of confidentiality). Before calculations start, we need to know the quarter revenue and the standard deviation of the 20 original observations. We also have to generate a random number from a standard normal distribution. This information is shown in the upper 3 rows of Table 13. Then, we calculated the average monthly revenue based on the quarterly data.

	REV_{t+1}	NAP_{t-2}	DAP_{t-2}	NAP_{t-3}	DAP_{t-3}	NAP_{t-4}	DAP_{t-4}	AOR_{t}	SOR_{t-}	ROR_{t}	AOR_{t-1}	SOR_{t-1}	ROR_{t-1}
REV_{t+1}	1.00												
NAP_{t-2}	0.22 (-0.09-0.50)	1.00											
DAP_{t-2}	0.32** (0.02-0.57)	0.95*** (0.93-0.98)	1.00										
NAP_{t-3}	0.18 (-0.14-0.46)	0.03 (-0.28-0.33)	0.05 (-0.26-0.35)	1.00									
DAP_{t-3}	0.29** (0.03-0.58)	0.12 (-0.19-0.41)	0.21 (-0.11-0.48)	0.96*** (0.92-0.98)	1.00								
NAP_{t-4}	0.01 (-0.30-0.31)	-0.07 (-0.37-0.24)	0.01 (-0.30-0.31)	0.01 (-0.30 -0.32)	0.06 (-0.25-0.36)	1.00							
DAP_{t-4}	0.15 (-0.16-0.44)	-0.01 (-0.31-0.30)	0.13 (-0.19-0.42)	0.09 (-0.23 -0.39)	0.21 (-0.11-0.48)	0.96*** (0.93-0.98)	1.00						
AOR_{t-1}	0.29* (-0.02-0.55)	0.13 (-0.18-0.42)	0.14 (-0.17-0.43)	0.21 (-0.11 -0.49)	0.23 (-0.08-0.50)	0.27* (-0.53-0.05)	-0.20 (-0.48-0.11)	1.00					
SOR_{t-1}	0.33** (0.03-0.58)	-0.01 (-0.31-0.30)	0.07 (-0.24-0.37)	0.13 (-0.18-0.42)	0.20 (-0.11-0.48)	-0.09 (-0.39-0.22)	0.02 (-0.29-0.32)	0.62*** (0.39-0.78)	1.00				
ROR_{t-1}	0.29* (-0.02-0.55)	-0.04 (-0.34-0.27)	0.02 (-0.29-0.32)	0.13 (-0.19-0.42)	0.18 (-0.14-0.46)	-0.12 (-0.41-0.20)	-0.02 (-0.33-0.29)	0.64*** (0.42-0.79)	0.99*** (0.97-0.99)	1.00			
AOR_{t-2}	0.21 (-0.10-0.49)	0.03 (-0.28-0.34)	0.08 (-0.24-0.38)	0.14 (-0.17-0.43)	0.15 (-0.17-0.43)	0.18 (-0.13-0.47)	0.20 (-0.11-0.48)	-0.06 (-0.36-0.25)	-0.10 (-0.39-0.22)	-0.11 (-0.40-0.21)	1.00		
SOR_{t-2}	0.33** (0.03-0.58)	0.05 (-0.35-0.26)	0.05 (-0.26-0.36)	0.00 (-0.30-0.31)	0.07 (-0.25-0.37)	0.11 (-0.21-0.40)	0.17 (-0.14-0.45)	0.17 (-0.15-0.45)	0.32** (0.02-0.57)	0.30* (-0.01-0.55)	0.62*** (0.39-0.78)	1.00	
ROR_{t-2}	0.32** (0.01-0.57)	-0.07 (-0.37-0.24)	0.02 (-0.29-0.32)	-0.03 (-0.33-0.28)	0.02 (-0.29-0.32)	0.10 (-0.22-0.39)	0.14 (-0.17-0.43)	0.16 (-0.16-0.44)	0.30 * (-0.01-0.55)	0.28* (-0.03-0.54)	0.64*** (0.42-0.79)	0.99*** (0.97-0.99)	1.00

Table 12. Correlation (95%-CI) between the variables for predicting revenue. * Correlation is significant at the 0.1 level, ** at the 0.05 level and *** at the 0.01 level.

To calculate an artificial standard deviation, we multiplied the random number with the standard deviation of the monthly revenue data. Adding this newly created standard deviation to the average revenue results in the first estimation of January 2020 revenue. We repeated this process 100 times, the average of which resulted in the final monthly revenue estimation. We followed a similar procedure for all other months from January 2020 till September 2021. This resulted in the adjusted data of which the descriptive statistics are shown in Table 11 (excluded because of confidentiality). Comparing the raw data with the adjusted data, the most striking change is the increase in sample size from 20 to 41 observations on all observations. For clarification, the independent variable data were extracted from a management information system and represent true values, differently than their corresponding revenue values.

Generally, both revenue and the independent variables have lower values in the adjusted data than the raw data. This can be explained by two reasons, the first being the impact of COVID-19. The adjusted data includes more months with COVID-19 measures, during which less patients had surgery of appointments at the outpatient clinic, which also resulted in a lower revenue. Additionally, costs have increased due to indexation over the past few years, and growth of OCON's patient population has also contributed to the increase in revenue. Therefore, data from a period longer ago has lower capacity and revenue values compared with more recent data. Also, the lowered standard deviation of revenue could be explained by the fact that outliers might be averaged out when we averaged 100 estimations per month to develop a final revenue estimation per month.

We assessed the presence of a linear relationship between revenue and the independent variables by using scatter and variable plots. We do not observe non-linear patterns in either of these plots. To check for the assumption of no perfect multicollinearity, we assess the correlations between the independent variables. Please see Table 12 for the correlation matrix. Like the correlations between the independent variables for predicting WIP, the number and duration of appointments at the outpatient clinic and the scheduled and realised OR-hours of the same months have high and significant correlations with narrow confidence intervals. To verify we do not violate the assumption of no perfect multicollinearity, we do not include these variables simultaneously in the forecasting model to not violate the assumption of linear regression and evaluate the VIF values after model construction.

Revenue quarter 1-2020		A
Standard deviation monthly revenue data		B
Random generated number from normal distribution Revenue quarter 1-2020		0.1374
1	Average revenue per month	A/3 = C
2	Artificial standard deviation	0.1374 * B = D
3	One estimation of revenue January 2020	C + D = E

Table 13. Example of generating monthly revenue values from quarterly data.

5.2.2 Raw data

Variable selection

We could not include all variables simultaneously in a model for variable selection, as this would result in zero residual degrees of freedom due to the small number of observations in the raw dataset. Therefore, we initially only included the capacity indicators in the explorative model and subsequently assessed the potential added value of the dummies using forward variable selection. This resulted in Model 3, which includes the variables number of appointments at the outpatient clinic three months ago, and available and scheduled OR-hours previous and current month. The results of Model 3 are presented in Table 14 (coefficients excluded because of confidentiality).

Besides the constant, all variables in Model 3 are significant. Following the same reasoning as in Section 5.1.2.2, we decided to not remove the constant from the equation despite it being insignificant. Different than Models 1 and 2, some estimates of Model 3 are counterintuitive. The coefficients for the variables available OR-hours previous month and scheduled hours current month indicate that revenue decreases when their value increases. This seems illogical as more used OR-capacity results in more opened DTCs that become revenue in the future. We assume that an increase of available OR-hours results in an increase in scheduled and realised surgeries as utilization rates have proven constant in the past.

On the other hand, the remaining variables are intuitive. They indicate that a rise in the number of scheduled appointments at the outpatient clinic, the available OR-hours current month, and the scheduled OR-hours previous month contributes to a revenue increase in the next month.

Model 3: $\boldsymbol{REV_{t+1} = \beta_0 + \beta_1\, NAP_{t-3} + \beta_2\, AOR_t + \beta_3\, AOR_{t-1} + \beta_4\, SOR_t + \beta_5\, SOR_{t-1} + \varepsilon_{t+1}}$

	Coefficient (Standard Error)	*T-value*
NAP_{t-3}		2.40**
AOR_t		3.32***
AOR_{t-1}		-3.26***
SOR_t		-3.13***
SOR_{t-1}		3.58***
Constant		0.64
Adjusted R^2	0.38	
RMSE		
F	5.14	
Sig.	0.03	

Table 14. Model for predicting revenue based on the raw data. Significance is indicated by asterisks, where *** indicates a significance at the 0.01 level, ** at the 0.05 level and * at the 0.1 level.

The fact that the model indicates that variables conveying the same information for different months result in opposite effects on revenue makes us question estimate reliability. For example, we do not expect the number of scheduled OR-hours in one month to negatively impact revenue and to positively impact revenue in another. This emphasizes once more that we should be cautious when using models based on small datasets.

Model performance

We use VIF values to detect the potential presence of multicollinearity in Model 3. These are shown in Table 15. Besides the NAP_{t-3} variable, all VIFs have values of around 10, suggesting that multicollinearity might impact our regression results. Multicollinearity does not reduce the explanatory power of the model but reduces the statistical significance of the independent variables. Removing either the available or scheduled OR-hours expectedly reduces multicollinearity, but also results in models without significant variables and predictive power, which is indicated by a negative adjusted R-squared.

The residual scatter plot for Model 3 is shown in Figure 15 (x-axis excluded because of confidentiality). We assume homoscedastic and uncorrelated error terms as we do not observe a pattern in the dots. This is also supported by the non-significant Breusch-Pagan test (p=0.17). Figure 16 shows the Q-Q plot (y-axis excluded because of confidentiality). As almost all points are scattered along the reference line, we assume that the error terms have a normal distribution. We contribute the observation outside the bandwidth to chance. We argue that the error term is independent since we included multiple capacity indicators relating to both the outpatient clinic and the OR in our model, and explored if adding dummies to correct for reduction and COVID periods improved forecasting performance.

We evaluate adjusted R^2 and RMSE to assess goodness of fit. Adjusted R^2 is 0.38 and considered to be significant because of the significant F-test. This value indicates that 38% of the variability in revenue is explained by Model 3. The mean, median, minimum, and maximum forecast accuracy of Model 3 are 91.4%, 93.2%, 73.6%, and 99.9%, respectively. Because of the sample size, we chose to not split the data into a test and training dataset. Therefore, we cannot assess the external validity of Model 3.

	VIF
NAP_{t-3}	1.23
AOR_t	10.85
AOR_{t-1}	9.65
SOR_t	10.63
SOR_{t-1}	10.42

Table 15. Variance inflation factors for Model 3.

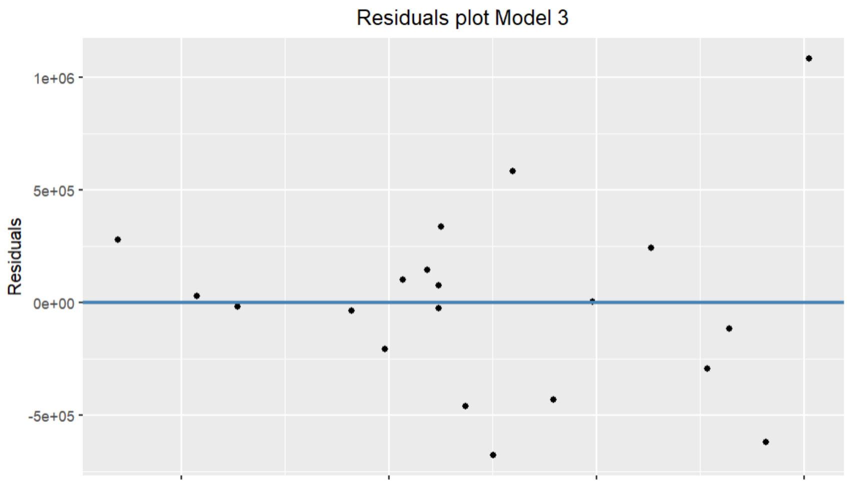

Figure 15. Residual scatter plot for Model 3.

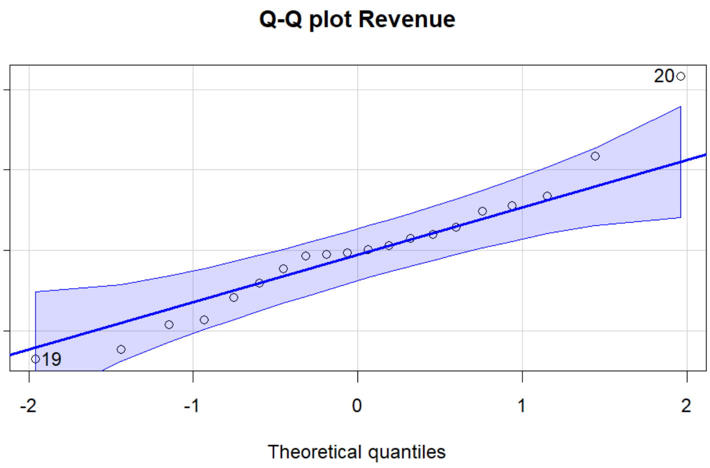

Figure 16. Q-Q plot for Model 3.

5.2.3 Adjusted data

We performed the same analyses as in Section 5.2.2 on the adjusted data, of which the results are shown in this section.

Variable selection

Because we more than doubled the size of our data set by generating monthly observations from the quartile data, we could include all variables simultaneously in a model for variable selection. This resulted in Model 4, which includes the variables number of appointments at the outpatient clinic three months ago, and available and scheduled OR-hours previous and current month. The results of Model 4 are presented in Table 16 (coefficients are excluded because of confidentiality).

All variables in Model 4 are significant, besides the "full" category of the reduction dummy and the constant. Again, we do not remove the constant from the model despite it being insignificant. Also, as described in Section 5.1.2.1, we cannot remove one category of a dummy variable whilst leaving another category in. Removal of the entire reduction variable results in a decrease of adjusted R^2 to 0.15 and the NAP_{t-2} and AOR_{t-1} variable losing its significance. As we want to have an as high as possible adjusted R^2 , and because full reduction occurs only twice per year, we chose to retain the dummy in the model.

Model 4: $REV_{t+1} = \beta_0 + \beta_1\, NAP_{t-2} + \beta_2\, SOR_t + \beta_3\, AOR_{t-1} + \beta_4\, REDUCTION(half)_{t-4} + \beta_5\, REDUCTION\,(full)_{t-4} + \varepsilon_{t+1}$

	Coefficient (Standard Error)	*T-value*
NAP_{t-2}		2.23**
SOR_t		3.08***
AOR_{t-1}		2.43**
$Reduction_{t-4}$ (half)		2.50**
$Reduction_{t-4}$ (full)		-0.06
Constant		-0.96
Adjusted R^2	0.25	
RMSE		
F	3.67	
Sig.	0.01	

Table 16. Model for predicting revenue on the adjusted data. Significance is indicated by asterisks, where *** indicates a significance at the 0.01 level, ** at the 0.05 level and * at the 0.1 level.

However, keeping a non-significant dummy in the model possibly impacts the estimates of the other independent variables, making them less reliable. Thus, there is no optimal solution regarding the in- or exclusion of the reduction dummy in Model 3. Also, the estimate of the dummy representing half reduction is not intuitive. In times of reduction, less patients are treated which is expected to negatively affect revenue, but the positive estimate in Table 16 suggests an opposite effect.

Model performance

Due to inclusion of the reduction dummy with more than two categories, we evaluate the assumption of no perfect multicollinearity by GVIF values. For Model 4, these are shown in Table 17. As the values are all close to one, we do not consider multicollinearity to impact our regression results.

The residual scatter plot for Model 4 is shown in Figure 17 (x-axis excluded because of confidentiality). Since we do not observe a pattern in the residuals, we assume homoscedastic and uncorrelated error terms. This is supported by the non-significant Breusch-Pagan test (p=0.49). In the Q-Q plot, which is shown in Figure 18, the dots are roughly scattered along the reference line (y-axis excluded because of confidentiality). Therefore, we consider normal distribution of the error terms. Observation 41 corresponds to observation 20 in Figure 16 and is attributed to chance. We argue that the error term is independent by following the same reasoning as provided for Model 3 in Section 5.2.2. For evaluating the goodness of fit of Model 4, we assess adjusted R^2 and RMSE. Adjusted R^2 is significant and has a value of 0.25. Mean, median, minimum, and maximum forecast accuracy are 90.7%, 93.1%, 66.0%, and 99.8%, respectively.

This minimum value corresponds to the revenue of May 2023, and can be explained as follows. Revenue was remarkably high this month because invoicing of DTCs started in 2023 was not possible before May. This inconvenience is a yearly recurring phenomenon caused by insurers that have not digitally agreed with the DTC prices yet.

	GVIF
NAP_{t-2}	1.13
SOR_t	1.08
AOR_{t-1}	1.17
$REDUCTION_{t-4}$	1.34

Table 17. Generalized variance inflation factors for Model 4.

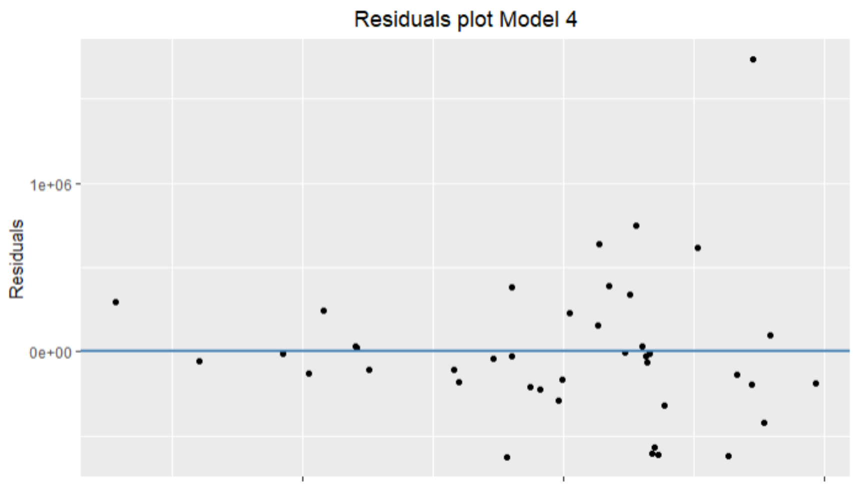

Figure 17. Residual scatter plot for Model 4.

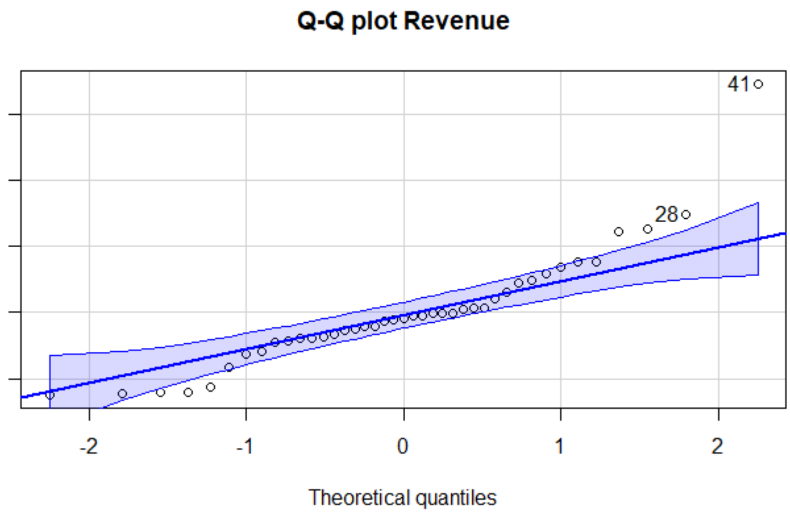

Figure 18. Q-Q plot for Model 4.

Although these prices have been agreed upon during the negotiations between insurers and providers, an additional digital check is required before invoicing can take place. As long as insurers have not done this, closed DTCs cannot be invoiced and remain to-be-invoiced. Once insurers have allowed invoicing, which usually happens in April or May, revenue is higher than would be expected based on used capacity. So, a forecasting model based on capacity indicators per definition underestimates revenue in such a situation.

For Model 4, we did not split the data into a test and training dataset to assess generalizability despite its size being larger than the dataset used for Model 3. One reason is that we wanted to be consistent for both approaches to forecasting revenue. A second reason is that we are uncertain about the reliability of the generated monthly revenue values. Although we used the standard deviation of the raw data and derived the monthly observations from the quarterly data, the generated monthly observations are still heavily influenced by chance. We cannot know if these values are representative of the true monthly revenues. Therefore, we should be careful in drawing conclusions.

Keeping that in mind, we can compare Models 3 and 4. Model 3 explains 13% more of the variability in revenue and has a RMSE that is 6.5% smaller than Model 4. Accuracy measures are comparable, with slight better performance of Model 3. Both models comprise intuitive and counterintuitive estimates. Model 4 includes two insignificant variables, whereas only includes one.

Section 6

Conclusion and discussion

In this section, we give an answer to our central research question based on our main findings. We describe the managerial implications accordingly. Furthermore, we discuss the strengths and limitations of our study, and conclude with providing some recommendations for further research.

NOTE: some text is excluded because of confidentiality. The confidential annex is registered in the repository of the University of Twente.

6.1 Conclusion

First, we interpret the main findings of Section 5 and use them to give an answer to our central research question. Subsequently, we summarize what our results mean in terms of action for OCON's financial management.

6.1.1 Main findings

Performing this study had a dual purpose. We wanted to gain insight into the factors constituting operating cash flows at OCON and their magnitude and timing, and enable adjustment of operational processes accordingly if desired. To do so, we aimed at answering the following central research question:

"How can OCON's cash flow from operations be forecasted?

For answering this question, our first approach was to use WIP as a proxy of operating cash flow. Results show that conservative WIP for the next month can best be forecasted by the number of appointments at the outpatient clinic in the current and previous month, when corrected for reduction and COVID-19 periods. Both appointment variables are positively correlated with conservative WIP, but model estimates indicate that the number of appointments in the current month has the biggest impact. In case of COVID-19 in the month we are predicting, WIP decreases with €.... Similarly, WIP is reduced with €… or €… in case of full and half reduction periods, respectively. It is intuitive that half reductions result in a smaller decrease in forecasted conservative WIP than full reductions, because larger or longer reductions leave less capacity to be used for seeing patients, thus producing WIP. The adjusted R^2 of 0.71 demonstrates good model fit, but performance decreases to 0.49 when the model is applied to new data, suggesting limited generalizability to new data. However, with a decrease of 0.9% compared with the training data, we still consider forecast accuracy on the test data set adequate with a median of 95.3%.

In descending order of positive impact on operative WIP, scheduled OR-hours in the next and current month have proven to be the best predictors when corrected for COVID. Interestingly, the reduction dummy did not have a significant effect on operative WIP, which contradicts our expectations. Since large reductions in the available OR-time, and thus scheduled OR-time, take place during holiday-periods, we expected this to be reflected by a significant reduction dummy in the model. The forecasting model for operative WIP seems to outperform the conservative one based on the explained variance of 82% but has the drawback of an insignificant constant. In our opinion, the disadvantages of excluding the constant outweighed the advantages, which is why it we kept it in the model. However, this does negatively affect model performance. Nevertheless, median forecast accuracy was 95.1% and 95.3% on the training and test set, respectively, and the percentage of explained variability and standard error of the residuals did not differ much between both datasets, indicating a robust model. This improved generalizability compared with the model for predicting conservative WIP might be explained by the larger number of degrees of freedom, which increases model parsimony and reduces sample size concerns.

We chose to split WIP and predict models for conservative and operative WIP separately to decrease the standard deviations of the model estimates and forecasts, thereby increasing accuracy. Therefore, to calculate total WIP, we should sum both forecasts. Consequently, this forecast is influenced by the standard deviations of two models rather than one. This means that the final forecast has a larger standard deviation than the separate models and therefore likely a smaller forecast accuracy.

Our second approach to answering the research question was to forecast revenue. Analyses on the raw data show that using the number of outpatient clinic appointments three months ago and the scheduled and available OR-hours current and previous month as predictors results in the best model with an adjusted R^2 of 0.38 and median forecast accuracy of 93.2%. Two of the five included independent variables, being available OR-hours previous month and scheduled OR-hours current month, show estimates that contradict our intuition and other model estimates. This, combined with the presence of multicollinearity in the models, makes us question estimate reliability.

The same goes for the revenue forecasting model based on the adjusted data, which includes the variables number of appointments two months ago, available OR-hours previous month, scheduled OR-hours current month, and reduction dummy four months ago as predictors. The dummy has a counterintuitive estimate, which suggests that revenue increases when there was a half reduction period four months ago. Based on goodness of fit, this model underperforms compared with the forecasting model on the raw data with an adjusted R^2 of 0.25. As the model based on the adjusted data has a mean forecast accuracy of 93.1%, both models seem similarly accurate in forecasting revenue.

Comparing all above-described results, we can assess which model is best suitable to forecast cash flows at OCON, thereby answering our research question. However, it is important to keep in mind that conclusions should be drawn with caution due to the small sample sizes. With small samples, statistical tests have limited power and results do not necessarily reflect the true values due to the high influence of chance and outliers on model results. This particularly applies to evaluating model robustness, as the test sets of the WIP forecasting models comprised only eight observations.

In our opinion, the models used for forecasting WIP outperform the revenue forecasting models based on the measures for goodness of fit, forecasting accuracy, logical variable estimates, and overall model significance. With F-test values of 60.23 and 16.95, we have greater evidence that the variables included in the models for forecasting WIP contribute to model fit than the ones included in the revenue forecasting models, as these have smaller F-test values of 5.14 and 3.67. Based on the adjusted R^2 cut-off point of 0.50, the WIP forecasting models are and the revenue forecasting models are not sufficiently capable to explain the variability observed in the revenue values. All forecasting models have good forecasting accuracy with percentages of over 90%. It should be noted that the accuracy measures for the revenue forecasting models might be an overestimation of reality as we only assessed forecast accuracy based on the data the models were build on. We do not know how the model will perform based on new data.

In the following section, we describe what these results imply for practice at OCON.

6.1.2 Managerial implications

Because of the limited data availability, we currently do not recommend applying the developed forecasting models in practice. This does not mean that the models by definition produce inaccurate forecasts, it means that we cannot know with a sufficient degree of certainty that the predicted financial outcomes adequately reflect reality. Based on this uncertainty, we do not recommend applying the model to enable informed decision making based on the results. In the future, more, quality data will become available. Adding these data will expectedly enhance model accuracy, reliability, and generalizability. Therefore, the results of this study serve as a blueprint for future forecasting models.

As rule of thumb, we recommend having a minimum of 30 observations for a model with one predictor and adding at least 10 observations for each additional independent variable. Also, when including a categorical variable, there should be 10 additional observations for the number of categories minus 1. So, for example 20 extra observations when including a dummy with three categories. Ideally, there would be enough available data to allow us to exclude the data from the COVID-19 pandemic. This would improve data quality as these observations are caused by an unusual situation and therefore do not reflect future values and complexify accurate cash slow forecasting.

On the other hand, the models could be used to estimate future WIP values as there currently is no alternative to forecast WIP. So, estimates generated by our models, although possibly inaccurate, do no harm as long as they are not used for important business decisions. This only applies to WIP, as these models have relatively more data and appear to have proper model performance. This does not apply to the revenue forecasts, as that dataset comprised only 20 raw data observations.

To forecast WIP, we should calculate conservative and operative WIP separately to subsequently add them up. We predict conservative WIP using the following equation:

$$\mathrm{WIP}_{\mathrm{c:}\,t+1} = \beta_0 + \beta_1 * \mathrm{NAP}_t + \beta_2 * \mathrm{NAP}_{t-1} - \beta_3 * \mathrm{COVID}_{t+1} - \beta_4 * \mathrm{COVID}_t - \beta_5 * \mathrm{Reduction}_{\,t+1}(\mathrm{full}) - \beta_6 * \mathrm{Reduction}_{t+1}\ (\mathrm{half})$$

Although we included the COVID-dummy in the above formula, it will not contribute to WIP forecasts in practice because its value will be zero. This is because the COVID-19 pandemic and its associated restrictions are over. In the months July and August, there is full reduction, meaning that we should fill out a 1 at the place of the $\mathrm{Reduction}_{t+1}$ variable when making forecasts for either of these two months. In the other months, we fill out a 0, which means that the variable does not contribute to the forecast. The half reduction variable should be used the same way, but for the months February, April, May, October, and December.

The other input variables, being number of appointments at the outpatient clinic of the current and previous month, can be extracted from business object (BO). In retrieving these values, it is important to ensure to only include appointments related to insured care. BO provides a filter to exclude the non-insure care appointments. Also, when making a forecast for the next month in the beginning of the current month, it could be the case that not all appointments for the current month are scheduled yet. Then, we should estimate this value by using a utilization rate. The average utilization rates per week and month are also shown in BO. However, utilization rates at the outpatient clinic have a relatively high variance. This means that the standard deviation of this predictor is increased when estimating the number of appointments with a utilization rate, which negatively impacts accuracy of conservative WIP forecasts.

For calculating operative WIP, we should follow a comparable approach, but use the following formula:

$$WIP_{o:\,t+1} = \beta_0 + \beta_1 * SOR_{t+1} + \beta_2 * SOR_t - \beta_3 * COVID_{t+1}$$

Again, the COVID-dummy in the above formula does not contribute to WIP forecasts in practice. The data regarding the scheduled OR-hours for the next and current month can be extracted from BO. We should use a utilization rate to correct for urgency surgeries, as these are only scheduled in the short-term. As utilization rates of the OR are quite constant throughout the year, estimates of OR-time are expected to be reliable. Also, urgent surgeries represent only a small portion of total OR-hours. The vast majority of OR-capacity is used for regular patients that are scheduled 12 to 6 weeks prior to their surgery date. To come to a final forecast of WIP, we should add up the calculated estimates of conservative and operative WIP.

Finally, we could use the forecasts of WIP to estimate TBI. However, these estimates are rather arbitrary as these are based on other estimations that include assumptions on for example utilization rates. Based on the WIP grouper extractions, average and median percentage of conservative WIP from the sum of conservative WIP and TBI, a rough estimation of conservative TBI can be generated. Similarly, an operative TBI estimation can be generated.

Operative TBI comprises a larger proportion of the total WIP and TBI, than its conservative counterpart does. This can be explained by the fact that operative DTCs have a shorter duration than conservative ones. Therefore, operative WIP and TBI have a smaller duration difference than conservative WIP and TBI. So, TBI is relatively longer included in the operative WIP dataset than the conservative one, explaining its higher percentage of the total number of DTCs included in the operative dataset.

6.2 Discussion

This final section serves as a discussion on this thesis. In the first part, we discuss some limitations and strengths of our study, followed by recommendations for further research.

6.2.1 Limitations

A major limitation of our study was the limited amount of available data and their quality. For forecasting WIP, we had a dataset comprising 53 observations of which 48 were left after data cleaning. This final dataset included data from the years 2020 and 2021, during which COVID-19 impacted the operations at OCON which resulted in valid, yet exceptional WIP values. To retain an as large as possible sample size, we chose not to delete the data observations heavily impacted by COVID-19 but included a dummy variable in the model. Although this was a good approach given the situation, it would have been preferable to exclude these data to improve model forecasting accuracy and generalizability if there was an abundance of data. That would also result the COVID-dummy to become superfluous in our forecasting model, because we do not expect a new COVID pandemic in the near future. Fewer included variables results in a simplified model, which is beneficial for statistical power, application in practice and interpretational purposes.

Statistical tests and validation techniques also have limited statistical power and internal and external validity in case of small sample sizes, because of the large influence of chance on the results. Also, outliers have a proportionally large effect on results in smaller datasets. In our study, this effect is magnified because of the use of RMSE as indicator of model performance. Although it is a commonly used measure of model fit, it should be noted that this measure may show worse model performance based on RMSE in case of outliers than if the model was fitted on data excluding outliers. This is caused by the fact that the calculation of RMSE includes squaring of the residuals, thereby increasing the relative influence of outliers, that have larger residuals than other data, on the overall results.

Additional statistical limitations include the fact that it was not possible to remove a non-significant intercept or non-significant category of a dummy variable from the models, without considerably changing the results or performance measures. Therefore, we were forced to include variables in our model that do not significantly differ from zero and thus not really add to model performance. This does not only complexify the models, but also influences the estimates of the remaining significant variables.

Besides underestimation of model performance due to outliers, it is also possible that we overestimated model performance because of the retrospectively collected data. Looking back, indicators regarding scheduled appointments or surgeries always match reality. However, when using the model for forecasting, we might not have this accurate information yet as schedules might not be completely filled. Also, rescheduling could lead to different numbers of scheduled appointments or surgeries at different moments in time. For example, we use the number of scheduled OR-hours of the next month as indicator of operative WIP. A month in advance, urgency cases are not scheduled yet. Therefore, a utilization rate is necessary to convert the scheduled OR-hours to a realistic value of future scheduled OR-hours. Because OR-utilization rates remain rather constant over time, using these would results in accurate estimations. This is more complicated for outpatient clinic utilization rates as these have more fluctuations. Although utilization rates can be an accurate tool for estimating capacity indicators, they are always outperformed by reality in retrospect. Therefore, quality of capacity indicators as predictors in the models could be an overestimation of reality.

Also, we assumed the start date of the DTC equals the date of the first appointment at the outpatient clinic, because the physician then opens the initial care path. This is true for most of the DTCs, as additional research, such as an X-ray, takes place approximately 30 minutes prior to the consultation with the physician. Sometimes however, a care activity, such as an MRI, is performed longer before the first appointment at the outpatient clinic. If this is the case, the start date of the DTC is backdated to the date of the first care activity. The difference in time between the first care activity and opening of the initial care path can vary from a few days to a few weeks. Because of this, the period between capacity use and the registration of the activities as WIP or revenue is shorter than the regular duration of the DTC as we presumed in our models. Although we do not expect this to have a substantial impact on our results because it applies to only a small percentage of conservative DTCs, it is important to be aware of it.

Finally, when transforming the raw revenue data to the adjusted data, we created monthly revenue observations from quartile data. This resulted in a dataset in which the sum of the three generated monthly values did not equal the original quartile value. Therefore, these monthly revenue values might not be a suitable representation of the true monthly values. This could be an explanation of the lower adjusted R^2 of the forecasting model on the adjusted data compared with the one on the raw revenue data.

6.2.2 Strengths

Besides the limitations, this study also has some strengths. The first one is that is concretizes the relationship between capacity and financial management at OCON. Although a connection was expected based on experience from practice and logical reasoning, this study demonstrates what part of capacity has which effect on the finances and with what timing. This contributes to a deeper understanding of value creation, as the entire process is considered from first appointment at the outpatient clinic until the receival of cash for the provided care activities. Also, we predicted finances in a broad sense as we focused on forecasting both short-term finances, being WIP, as well as longer term, being revenue.

As stated before, the results of our study mainly serve as a blueprint and should be enhanced by the addition of new data in the future. That is why we aimed at developing a robust model that better handles new data, for example by not overly perfectionating the data during the data cleaning phase and by including a dummy variable for reduction periods. Also, we split WIP into a conservative and operative part and developed separate models for both to decrease the standard deviations of the forecasts and increase their accuracy.

Furthermore, we already had insight into the patient and scheduling processes, because of the master thesis Health Sciences I have already finished at OCON. This provided me with information about the business operations that are relevant for deeper understanding of the connection between capacity and finances.

6.2.3 Recommendations for further research

This study serves as starting point for forecasting operating cash flows at OCON. In the future, additional studies can build on to our results to improve forecasts or explore other parts of operating cash flows. We will provide some recommendations for future research below.

Firstly, as was already mentioned several times before, model accuracy, reliability, and generalizability will increase when sample size increases. Larger sample sizes might also allow for the deletion of data observations caused by an exceptional situation, such as COVID-19, thereby improving model fit. Therefore, our study can be regarded as a guideline for a future researcher to perform a similar study but with additional data. To better assess the generalizability of the developed models, different validation approaches could also be applied, such as K-fold and Leave-One-Out cross-validation. Different than the validation set approach in which a different split could result in a considerably different results, these techniques mitigate the impact of the random split of the dataset on the external validity assessment.

Additionally, extra variables could be explored as predictors of WIP or revenue, such as the available outpatient clinic capacity. We did not include this variable in our models, because this was only registered in Business Objects from October 2021 on, and its values were incorrect until November 2022 due to registration errors in the outpatient clinic schedules. Currently, this information is accurately recorded, indicating that its potential value as predictor can be assessed in future forecasting models. We cannot estimate the magnitude and significance of this variable, but we expect it to be a convenient predictor in practice for several reasons. The first is that the available outpatient clinic is determined several months in advance and is not susceptible to changes. Also, this variable is not dependent on the degree to which the schedules are already filled and therefore does not require calculations with utilization rates. So, we consider it worthwhile to evaluate the predictive ability of this variable in the future.

Besides investigating the predictive ability of additional variables, forecasting models for longer forecasting horizons could also be explored. In this study, we focused on predicting WIP and revenue at t+1. Future research could focus on predicting periods further into the future, being for example t+2 or t+3. We expect these models to result in less accurate predictions because of the increased length of time into the future for which forecasts are made, which decreases forecast reliability.

Furthermore, we excluded TBI in our forecasts because WIP seemed to have a stronger correlation with the capacity indicators. However, TBI still represents approximately 20 percent of total WIP and TBI. Currently, the only possibility to estimate TBI is based on a ratio with WIP as suggested in Section 6.1.2, but this will most likely not result in accurate estimates. That is why we recommend future research to focus on developing a forecasting model that also or only includes TBI. Using the sum of WIP and TBI as outcome indicator instead of only WIP could be a possibility, but this could cloud the relation between WIP and the capacity indicators, and result in a model with lowered model fit and larger standard deviations. Therefore, predictor variables for TBI should also be included. These could be identified by mapping and analysing the process money follows over time until it is transformed to an invoice. Examples of TBI variables could relate to the days the grouper and invoicing runs are made, the number of runs per month, the percentage of faulty, non-claimable care product extractions the grouper, and the duration between invoicing run and the sending of the invoice.

Finally, to improve the revenue forecasts, a model could be developed in which revenue is corrected for the influential observations. As described in Section 5.2.1, these influential observations are caused by the fact that invoicing is not possible because the insurers have not digitally approved the new DTC prices yet. When including this issue in the model, for example by using a dummy variable, we expect model performance to increase.

References

Ball, R., & Nikolaev, V. V. (2022). On earnings and cash flows as predictors of future cash flows. *Journal of Accounting and Economics, 73*(1), 101430.

Boot, J. M. (2013). *De Nederlandse gezondheidszorg* (9th ed.). Zeist: Bohn Stafleu van Loghum.

Coyne, J. S., & Singh, S. G. (2008). The early indicators of financial failure: A study of bankrupt and solvent health systems. *Journal of Healthcare Management, 53*(5), 333-345. doi:10.1097/00115514-200809000-00010

Drtina, R. E., & Largay III, J. A. (1985). Pitfalls in calculating cash flow from operations. *Accounting Review*, 314-326.

El-Sayed Ebaid, I. (2011). Accruals and the prediction of future cash flows: empirical evidence from an emerging market. *Management research review, 34*(7), 838-853.

Fatemeh, S. (2020). Cash flow forecasting by using simple and sophisticated models in Iranian companies*3*(1), 24-52. Retrieved from doi:10.22034/ijf.2020.202650.1071

Habib, A. (2010). Prediction of operating cash flows: Further evidence from Australia. *Australian Accounting Review, 20*(2), 134-143.

Holmes, G. M., Kaufman, B. G., & Pink, G. H. (2017). Predicting Financial Distress and Closure in Rural Hospitals. *The Journal of Rural Health, 33*(3), 239-249. doi:https://doi.org/10.1111/jrh.12187

Jemaa, O. B., Toukabri, M., & Jilani, F. (2015). Accruals and the prediction of future operating cash-flows: Evidence from Tunisian companies. *International Journal of Accounting and Economics Studies, 3*(1), 1-6.

Jeurissen, P., & Maarse, H. (2021). *The market reform in Dutch health care: results, lessons and prospects*.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLoS One, 13*(3), e0194889.

McLaney, E., & Atrill, P. (2018). *Accounting and Finance: An Introduction* (9th ed.). Harlow: Person EducationLimited.

Minister of Health Welfare and Sport J.F. Hoogervorst. (2003). *Invoering Diagnose Behandel Combinaties (DBCs); Brief minister over de invoering van een nieuwe bekostigingssystematiek in de ziekenhuiszorg* (29248 (nr. 1)). 's-Gravenhage: Sdu Uitgevers Retrieved from https://www.parlementairemonitor.nl/9353000/1/j9vvij5epmj1ey0/vi3aljzdq4zu.

Mulenga, M., & Bhatia, M. (2017). The review of literature on the role of earnings, cash flows and accruals in predicting of future cash flows. *Accounting and finance Research, 6*(2), 59-70.

Nederlandse Zorgautoriteit. (2022). *Handleiding dbc-systematiek*. (PUC_711001_22). Utrecht Retrieved from https://puc.overheid.nl/doc/PUC_711001_22.

Nederlandse Zorgautoriteit. (2023). *Regeling medisch-specialistische zorg- NR/REG-2306a*. Nederlandse Zorgautoriteit.

Park, E., Chae, B., & Kwon, J. (2018). Toward understanding the topical structure of hospitality literature: Applying machine learning and traditional statistics. *International Journal of Contemporary Hospitality Management, 30*(11), 3386-3411.

Takhtaei, N., & Karimi, H. (2013). Relative ability of earnings data and cash flow in predicting future cash flows. *International Journal of Accounting and Financial Reporting, 3*(1), 214-226.

Tangsucheeva, R., & Prabhu, V. (2014). Stochastic financial analytics for cash flow forecasting. *International Journal of Production Economics, 158*, 65-76. doi:10.1016/j.ijpe.2014.07.019

Tweede Kamer der Staten-Generaal. (2004). *Regeling van een sociale verzekering voor geneeskundige zorg ten behoeve van de gehele bevolking (Zorgverzekeringswet)*. (29763). 's-Gravenhage: Sdu Uitgevers Retrieved from https://zoek.officielebekendmakingen.nl/kst-29763-3.html.

UCLA: Statistical Consulting Group. (2023). FAQ: WHY ARE R2 AND F SO LARGE FOR MODELS WITHOUT A CONSTANT? Retrieved from https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-why-are-r2-and-f-so-large-for-models-without-a-constant/

van der Spoel, S., van Keulen, M., & Amrit, C. (2013). *Process Prediction in Noisy Data Sets: A Case Study in a Dutch Hospital.* Paper presented at the Data-Driven Process Discovery and Analysis, Berlin, Heidelberg.

World Health Organisation. (2023). International Statistical Classification of Diseases and Related Health Problems (ICD). Retrieved from https://www.who.int/standards/classifications/classification-of-diseases

Wu, W., & Zhou, Z. (2021). A comprehensive way to access hospital death prediction model for acute mesenteric ischemia: A combination of traditional statistics and machine learning. *International Journal of General Medicine, 14*, 591-602. doi:10.2147/IJGM.S300492

Zaniletti, I., Larson, D. R., Lewallen, D. G., Berry, D. J., & Maradit Kremers, H. (2023). How to Develop and Validate Prediction Models for Orthopedic Outcomes. *Journal of Arthroplasty, 38*(4), 627-633. doi:10.1016/j.arth.2022.12.032

Appendices

Appendices are excluded because of confidentiality.