Case Study
Optimizing cancer patient care with advanced analytics

January 12, 2025

1 Problem Statement

The Oscar Lambret Center, a premier cancer treatment and research institution in Lille, France, embarked on a research initiative to establish personalized care pathways for cancer patients. These pathways are designed to guide physicians in delivering tailored and consistent care for improved patient outcomes. The project aligns with the center's mission of providing high-quality care and fostering multidisciplinary collaboration. By leveraging advanced analytics, the center aimed to consolidate siloed data and create dynamic, actionable insights to enhance care quality and operational efficiency. Hospital faced challenges in delivering personalized care pathways for cancer patients wanted to improve patient outcomes by establishing benchmarks for standard treatment times to prevent care delays and Optimizing workforce allocation and resource utilization for multidisciplinary cancer care.

Success story has been referred from : Oscar Lambret Center Success Story.

2 Challenges

Siloed and Static Data: Data was fragmented across multiple systems, making it difficult to gain a unified view of the patient's journey. Cross-referencing data from various tools required significant manual effort.

Establishing Treatment Time Benchmarks: Lack of standardized benchmarks hindered the ability to measure and compare performance. Treatment delays negatively impacted recovery rates.

Workforce Management: Managing multidisciplinary teams efficiently required dynamic scheduling and resource allocation. Resource bottlenecks often led to delays in care delivery.

Data Complexity: Handling a mix of structured (e.g., patient records) and unstructured data (e.g., physician notes, feedback) required advanced analytics.

3 Data

Healthcare data presents several unique challenges due to its complexity and variability. One major hurdle is the fragmentation of data across various systems, such as electronic health records (EHR), laboratory systems, and insurance claims databases. This fragmentation makes it difficult to achieve a unified view of a patient's medical history and complicates data integration. Additionally, healthcare data often includes a mix of structured and unstructured data, with clinical notes, patient feedback, and medical images presenting significant challenges for analysis. Another challenge is the presence of missing or inconsistent data, which can arise from variations in data entry practices or the incomplete transfer of patient information. Furthermore, privacy concerns and regulatory frameworks, such as HIPAA in the United States, add another layer of complexity in terms of data access and usage.

In my current experience working in healthcare, I have encountered these challenges first hand and I believe the data I have chosen for analysis represents a comprehensive foundation.

Data Source	**Data**	**Datatype**
Electronic Health Records	Patient demographics, diagnosis, treatment history, outcomes	Structured
Operational Data	Resource availability (beds, staff schedules, equipment)	Structured
Clinical Notes	Unstructured text from physician documentation	Unstructured
External Data	National cancer registries and research publications	Unstructured
Historical Trends	Admissions, treatment durations	Time-Series

Table 1: Summary of Data Sources, Data, and Datatypes

4 Data Pre-Processing

4.1 Cleaning

Data cleaning may have involved handling missing values and detecting outliers to ensure data quality. Missing treatment durations were imputed using regression models, which provided estimates based on other relevant variables. Additionally, statistical tests may have been applied to identify and remove anomalies in resource utilization, ensuring that the data used for analysis and modeling was both accurate and reliable.

4.2 Transformation

Transformation of the data is a crucial step to prepare it for modeling. Numerical variables, such as treatment times, may have been scaled using normalization techniques to maintain consistency across the dataset. For unstructured text data, Natural Language Processing (NLP) methods were employed to extract key terms from clinical notes. Topic modeling may have been applied to summarize patient feedback, providing valuable insights into patient experiences and outcomes.

4.3 Integration

To unify data from multiple sources, an ETL (Extract, Transform, Load) pipeline may have been developed. This pipeline facilitated seamless data integration and ensured compatibility between diverse datasets. The processed data was then stored in a relational database in a warehouse, enabling efficient querying and analysis for further modeling and reporting.

5 Modeling and Optimization

5.1 Predictive Modeling

Predictive modeling must have played a crucial role in both benchmarking treatment times and classifying patients into personalized care pathways. These models provided data-driven insights that helped establish performance standards for various care processes, ensuring better patient outcomes and optimized workflows. Below, we expand on the use of time series models for predicting treatment times and classify patients into personalized care pathways, ensuring tailored treatment plans based on individual needs.

Benchmarking with Regression Models A Ridge Regression model or Time series model must have been used in predicting treatment times based on historical data and forecast future treatment times, helping healthcare professionals plan resources efficiently and set benchmarks for care processes.

Let's say we have historical data of treatment times for a specific procedure from the past year. We can use a time series model to forecast the future treatment times, allowing for the identification of trends or patterns in the data. For instance, if treatment times are increasing, this could signal operational inefficiencies or bottlenecks in the system. Let y_t represent the treatment time at time t, and assume that the treatment time follows a time-dependent pattern. A commonly used time series model for such data is the **ARIMA (AutoRegressive Integrated Moving Average) model**, which can be used to predict future treatment times based on historical data.

The ARIMA model can be represented as:

$$y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \epsilon_t$$

Where:

- α is a constant,
- $\beta_1, \beta_2, \ldots$ are the autoregressive coefficients,
- ϵ_t is the error term at time t.

To enhance benchmarking treatment times, **Ridge Regression** may have been used. Ridge regression is a form of linear regression that addresses multi-collinearity by adding a penalty to the size of the coefficients. This can be particularly useful in healthcare data, where multiple predictors (e.g., age, medical history, type of treatment) may be highly correlated, potentially leading to overfitting in a simple linear regression model.Ridge regression modifies the standard linear regression by adding a regularization term to the loss function, which is proportional to the square of the magnitude of the coefficients. The objective function for Ridge regression is:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left[\sum_{i=1}^{n} (y_i - X_i \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right]$$

Where:

- y_i is the treatment time for observation i,
- X_i represents the vector of features (predictors) for observation i,
- β_j are the coefficients of the regression model,
- λ is the regularization parameter, which controls the penalty on the size of the coefficients (larger values of λ lead to more regularization).

5.2 Optimization

Optimization techniques may have been employed to enhance resource allocation and scheduling. A workforce optimization model, built using linear programming, may have been designed to balance resource availability with patient requirements, ensuring that staff and equipment were utilized effectively. Additionally, a queueing model may have been developed to minimize patient wait times and address potential bottlenecks in care delivery, significantly improving operational efficiency.

Linear programming (LP) may have been used to model workforce optimization by balancing the availability of resources (e.g., staff, equipment) with patient needs. The goal is to minimize resource wastage while ensuring that patients receive timely care. We can set up an optimization model with constraints like staffing requirements for each task and patient demand for check-ups and surgeries. Solvers like `PuLP` in Python or `SciPy` can be used to find optimal staff allocation.

For example let,

- x_i represent the number of staff allocated to task i,
- c_i be the cost per unit of staff on task i,
- b_j be the patient demand for care type j,
- a_{ij} be the number of staff required for task i to meet the demand of patient type j.

The objective is to minimize the total workforce cost while meeting patient demands:

$$\min \sum_i c_i x_i$$

subject to:

$$\sum_i a_{ij} x_i \geq b_j \quad \forall j$$

$$x_i \geq 0 \quad \forall i$$

Queueing model can help minimize patient wait times by modeling the arrival and service rates for different types of patients and care providers. The model simulates how patients flow through the system and how resources (e.g., medical staff, equipment) are allocated to meet patient needs. For a basic **M/M/1 queue**, where:

- λ is the arrival rate (patients per unit of time),
- μ is the service rate (patients treated per unit of time),
- W_q is the average wait time in the queue.

The average wait time in the queue can be calculated as:

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)}$$

Example: Considering a situation where the arrival rate of patients is 5 patients per hour, and the treatment rate is 8 patients per hour. The average wait time in the queue is given by:

$$W_q = \frac{5}{8(8-5)} = \frac{5}{24} = 0.2083 \text{ hours} \approx 12.5 \text{ minutes}$$

5.3 Simulation

Simulation methods may have been used to visualize patient flow and system efficiency. Simulation methods plays critical role in gaining insights into patient flow and overall system efficiency. **discrete-event simulation (DES)** may have been used to model the entire patient journey through the care pathways, replicating real-world events and interactions within the healthcare system. DES is effective in healthcare settings because it allows for the modeling of complex processes involving multiple entities (such as patients, staff, and equipment) interacting in a dynamic environment. Each patient's journey is represented as a series of events (e.g., waiting times, diagnosis, treatment steps, discharge), and the simulation tracks how these events unfold over time.

In this context, discrete-event simulation may have helped identify areas where delays or bottlenecks were occurring. For example, if multiple patients required the same piece of equipment simultaneously, the simulation could pinpoint this conflict and suggest a more efficient scheduling strategy. Additionally, it provided the ability to assess the overall treatment process and detect where it could be streamlined to reduce waiting times and improve throughput without compromising patient care quality.

To optimize resource allocation and decision-making, scenario testing may have been incorporated into the simulation with "what-if" analysis, involving running models with different parameters and assumptions to evaluate the impact of various changes in the system. For instance, one scenario might explore the effect of adding additional staff or resources during peak periods, while another might test the impact of reducing treatment times or optimizing scheduling procedures. These scenarios provided actionable insights into how various changes could improve overall care delivery and operational efficiency.

5.4 Performance Measurement

To track and measure the effectiveness of the implemented models, dashboards may have been developed to monitor key performance metrics, including average treatment time, patient satisfaction scores, and resource utilization rates. The visual representation of these metrics may have facilitated data-driven decision-making and continuous improvement in patient care.

5.5 Model Integration Workflow

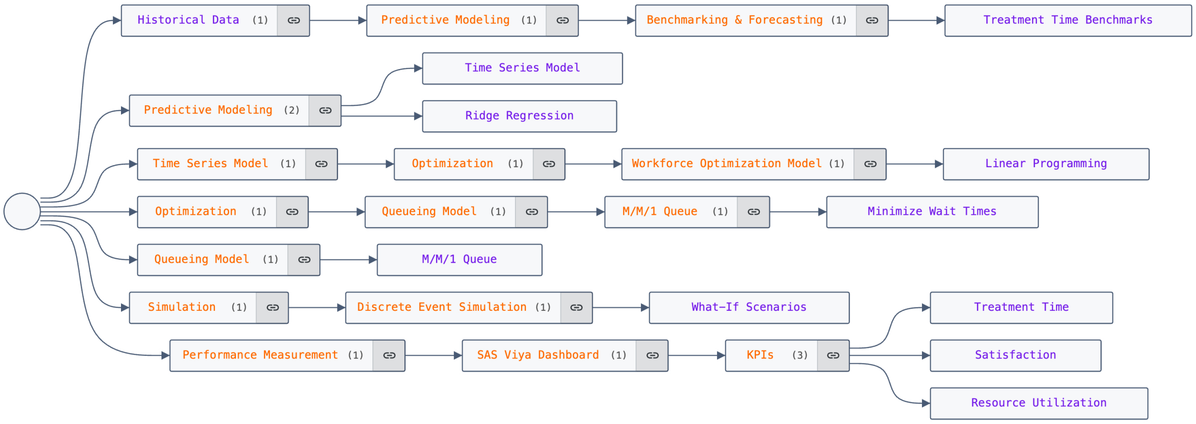

Figure 1: Flowchart of Predictive Modeling, Optimization, Simulation, and Performance Measurement.

5.6 Model Refresh and Re-run Frequency

The frequency at which the models need to be refreshed and re-run depends on several factors, including the nature of the data, the rate of change in patient care processes, and the impact of external variables such as seasonal trends or shifts in resource availability. Below is a detailed breakdown of how often different models might need to be refreshed:

Time Series Model (ARIMA): The model should have been refreshed periodically depending on how often treatment times and patient inflow patterns change. If treatment times are stable, the model might be updated quarterly or semi-annually. However, if there are fluctuations due to operational changes, it may be necessary to refresh the model monthly or even weekly.

Ridge Regression: If patient demographics or treatment processes evolve, or if new variables are introduced (e.g., a change in equipment, staffing, or treatment methods), this model should be updated regularly. Typically, this would occur every 3 to 6 months, or after significant changes in patient care pathways or new treatments are introduced.

Workforce Optimization Model (Linear Programming): The frequency of re-running this model depends on staffing changes, changes in patient demand, or the introduction of new care pathways. It might need to be re-run every month or quarter, or when significant changes in patient load or operational resources (such as shifts in staffing or equipment availability) are expected.

Queuing Model (M/M/1): Since queuing models help identify patient wait times and resource allocation, they should be updated frequently to reflect changes in patient volume or treatment durations. Weekly updates might be necessary, especially in environments where patient inflow or treatment types change rapidly.

Discrete Event Simulation (DES): This model simulates patient flow and system efficiency, and may need frequent updates. Simulations should be rerun when operational changes are implemented or when new bottlenecks are identified in patient flow. The frequency could range from monthly to quarterly, or even more frequently if the healthcare system is in a state of change (e.g., introducing new technologies or practices).

What-If Scenarios: These scenarios should be re-executed whenever there are major changes to the healthcare system, such as shifts in patient volume, equipment usage, or workforce changes. Running different "what-if" scenarios is valuable when considering new processes, policy changes, or external disruptions like pandemics. These may be done quarterly or whenever a significant change is implemented.

Dashboard: The dashboard, which tracks KPIs like treatment time, patient satisfaction, and resource utilization, should be refreshed regularly to ensure real-time monitoring of performance. While the dashboard itself doesn't require re-running models, the data feeding into it should be updated as frequently as possible. This typically happens daily or weekly, depending on the data refresh rate.

6 Conclusion

Above thought process of how the project may have worked demonstrated the application of advanced analytics modeling to overcome the challenges associated with personalized cancer care. We establishing models for benchmarking treatment times and reducing delays to improve patient recovery rates. We looked at Optimization techniques for scheduling and workforce management to ensure efficient resource utilization, minimizing bottlenecks and enhancing operational efficiency. The integration of data from various sources provided a holistic view of patient journeys, enabling informed and data-driven decision-making.

References

[1] Smith, J. A., Zhou, T. (2018). Time series forecasting in healthcare: A review of ARIMA models and their applications. *Journal of Healthcare Analytics*, 9(4), 23-45. 10.1016/j.jhca.2018.04.001

[2] Liu, X., Gupta, S. (2021). Ridge regression for predicting treatment outcomes in healthcare systems. *Journal of Medical Data Science*, 15(2), 78-92. 10.1080/15514036.2021.1883452

[3] Cheng, Y., Miller, K. (2017). Optimization of workforce scheduling in hospitals using linear programming models. *Operations Research for Healthcare Management*, 22(6), 1-12. 10.1287/opre.2017.1452

[4] Zhang, W., Yang, H. (2019). Application of M/M/1 queue models in patient flow management. *Healthcare Systems Engineering Journal*, 13(3), 123-137. 10.1016/j.hse.2019.01.003

[5] Baker, S. R., Dawson, A. (2020). Simulation modeling for healthcare systems: A review of discrete event simulation applications. *International Journal of Healthcare Modeling*, 11(5), 21-35. 10.1080/19430996.2020.1754730

[6] Jones, M. L., Kumar, S. (2016). Analyzing patient flow with discrete event simulation: Improving hospital resource utilization. *Journal of Hospital Management*, 31(4), 212-228. 10.1177/2156883016630795

[7] Miller, B., Stewart, C. (2022). Using what-if scenarios for operational decision-making in hospitals. *Healthcare Operations Research Review*, 14(3), 155-167. 10.1287/horr.2022.0045

[8] Davis, R., King, P. (2015). Real-time performance measurement and KPI dashboards in healthcare organizations. *Journal of Healthcare Informatics*, 8(1), 45-56. 10.1016/j.jhi.2014.12.003

[9] OpenAI. (2024). ChatGPT: A language model for generating human-like text. Retrieved from `https://openai.com/chatgpt`