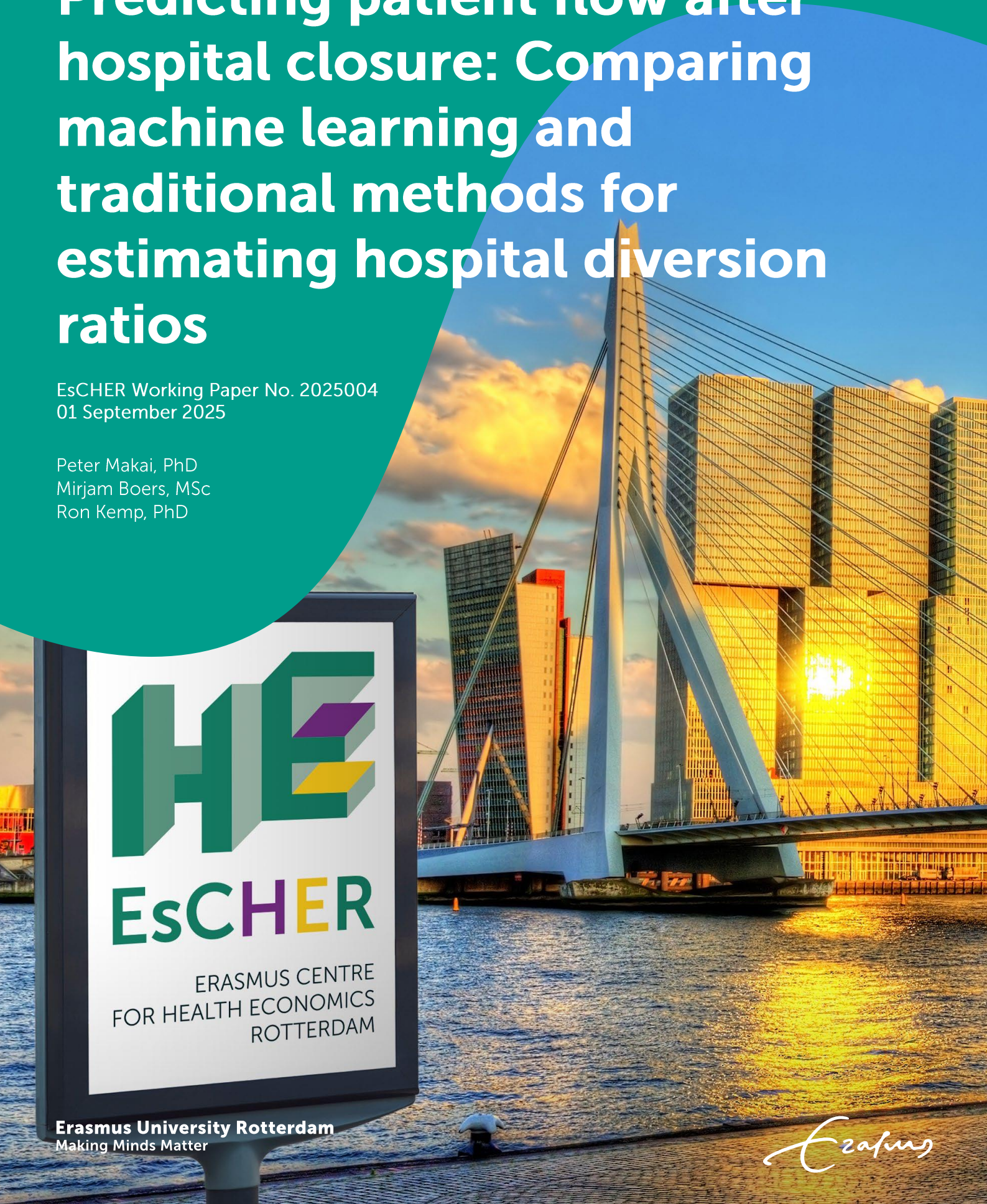
Predicting patient flow after hospital closure: Comparing machine learning and traditional methods for estimating hospital diversion ratios

EsCHER Working Paper No. 2025004
01 September 2025

Peter Makai, PhD
Mirjam Boers, MSc
Ron Kemp, PhD



Erasmus University Rotterdam
Making Minds Matter

Predicting patient flow after hospital closure: Comparing machine learning and traditional methods for estimating hospital diversion ratios

Authors
Peter Makai PhD Erasmus University Rotterdam
Mirjam B. Boers, MSc
Ron Kemp PhD, Authority for Consumers and Markets, Erasmus University Rotterdam

Corresponding author and contact details: Peter Makai makai@eshpm.eur.nl

Keywords
Hospital merger control, health service planning, prediction models, diversion ratios

JEL classification
I11,I18,L40,C53,C55

Cite as
Makai,P.,Boers,MB.,Kemp,R. 2025. Predicting patient flow after hospital closure: Comparing machine learning and traditional methods for estimating hospital diversion ratios. EsCHER Working Paper Series No. 2025004, Erasmus University Rotterdam. Available from: https://www.eur.nl/en/research/escher/research/working-papers

Erasmus Centre for Health Economics Rotterdam (EsCHER) is part of Erasmus University Rotterdam.
Want to know more about EsCHER? Visit www.eur.nl/escher
Want to contact EsCHER? E-mail escher@eur.nl

Interested in more EsCHER Working Papers? Download from www.eur.nl/escher/research/workingpapers

Abstract

Accurately predicting patient flows is crucial for applications in merger control and health service planning. In this study, we compared six methods of predicting the change in patient flows (diversion ratios) after hospital closure, using the case of the 2018 bankruptcy of the Slotervaart Hospital in Amsterdam, The Netherlands. We focused on three patient groups: cataract, ear, nose, and throat (ENT), and intestinal cancer. We used the pre-bankruptcy period January 2016—June 2018 as the development set as the hospital had a stable patient inflow in this period. The post-bankruptcy period January 2019—December 2020 was used as the test set, when the bankruptcy was finalized. To avoid endogeneity caused by the patient following their specialist, we restricted our analysis to patients newly referred to the hospital by their general practitioner. We compared predictive performance of 1) patient flow analysis based on an allocation proportional to the market shares, 2) conditional logit, 3) mixed logit, 4) least absolute shrinkage and selection operator (LASSO), 5) random forest (RF), and 6) gradient boosting machine (GBM) based on mean absolute difference (MAD) and on improvement in root mean square error (RMSE). Predictive performance was the highest for RF (mean MAD across all hospitals ranging from 3.67%-4.08% per patient group, mean RMSE ranging from 0.97%-1.29%). GBM followed, with mean MAD ranging from 3.58% to 6.00% and mean RMSE from 0.92% to 1.90%. LASSO showed mean MAD of 20%–44% and mean RMSE of 6.2%–14.7%. Conditional logit had mean MAD of 22%–44% and mean RMSE of 5.7%–14.7%, while mixed logit performed similarly with mean MAD of 22%–44% and mean RMSE of 5.7%–14.78%. Patient flow analysis was the least accurate, with mean MAD of 25%–48% and mean RMSE of 7.48%–17.65%. Given its consistently low error rates, RF appears to be the most promising method for improving the accuracy of merger and centralization decisions.

Introduction

Diversion ratios are a critical input for both hospital merger enforcement (Capps, Dranove, & Satterthwaite, 2003) and health services planning (Aggarwal et al., 2022). They measure the extent to which consumers will substitute from one product or service to another in case of a price rise or the withdrawal of a supplier from the market. In the hospital merger context, a high diversion ratio between two hospitals means that patients are more willing to switch between them; that is, higher diversion ratios indicate that the hospitals are closer substitutes. When negotiating with a hospital that has a closely substitutable competitor hospital, a health insurer will generally be able to negotiate more favorable prices because, if the hospital demands too high a price, the insurer can credibly threaten to shift patients to the competitor hospital (Garmon, 2017). Consequently, mergers involving closely substitutable hospitals raise greater competitive concerns. Diversion ratios provide a measure of that substitutability (Garmon, 2017). In a health services planning context, diversion ratios can inform the allocation of physicians or infrastructure (Aggarwal et al., 2022).

In both settings, conclusions and policy implications will be more accurate as diversions are more reliably estimated. Historically, diversions were commonly estimated through patient flow analysis using market shares based on postcode-level data (Heida, van Engelsen, Baeten, & van Gent, 2016) and choice models, which typically included distance between the patient's home and the hospital as the primary explanatory variable (van der Geest & Varkevisser, 2024). Over the last decade, researchers developed machine learning methods that can also provide estimates of diversions (Raval, Rosenbaum, & Wilson, 2022). In this paper, we evaluate the predictive performance of six methods of computing diversion ratios: 1) patient flow analysis based on historical market shares, 2) conditional logit, 3) least absolute shrinkage and selection operator (LASSO), 4) mixed logit, 5) random forest (RF), 6) gradient boosting machine (GBM). We focus on a specific event, the closure of Slotervaart Hospital in Amsterdam, due to bankruptcy on October 25th 2018. For each of the six approaches, we use the same pre-closure data to estimate diversion ratios and use those to generate predictions of post-closure patient flows. To assess the predictive performance of each estimation method, we then compare these predicted patient flows to the observed post-closure patient flows originating from the 4 level postcode areas. We do this for three groups of patients: cataract, ear, nose and throat (ENT), and intestinal cancer.

We find that machine learning methods, particularly RF are much more accurate with lower mean absolute difference (MAD) as well as lower root mean square error (RMSE), than patient flow analysis, conditional logit, LASSO and mixed logit.

Our study builds on the analysis by (Rossi, Whitehouse, & Moore, 2018) who conducted a related patient flow study based on GP-referrals and compared GP-referral analysis to choice models. In a GP-referral analysis, historical GP-referrals are used to obtain market shares, based on the number of patients referred by GPs to each hospital. This analysis was compared with choice models, yielding similar results. Due to the minor differences between the two approaches, (Rossi, Whitehouse, & Moore, 2018) recommend continuing the use of patient flow analysis based on market shares in practice. In addition, our work also relates to (Raval, Rosenbaum, & Wilson, 2021) (Raval,

Rosenbaum, & Wilson, 2022) who used disasters such as hurricanes or earthquakes as exogenous shocks to hospital patient preferences, when comparing the performance of various machine learning and choice models to predict diversion ratios. Specifically, (Raval, Rosenbaum, & Wilson, 2021) studied several disaster scenarios. They found that machine learning methods are more accurate in general, while showing the limits of the methods when the patient choice set is severely impacted by disasters (Raval, Rosenbaum, & Wilson, 2021). Although disasters are truly exogenous, they often force residents to relocate temporarily, altering their healthcare choices. The choice set of these fleeing patients therefore changes not only due to the hospitals that close due to the natural disaster, but also due to living at a different address. Relying on permanent addresses in administrative data introduces bias in distance-based models, as it misrepresents patients' actual locations during the disaster. Therefore, such analyses using disaster-data are less useful for competition policy or health system planning, which must address the counterfactual scenario where only the hospital is removed from the choice set, without other disruptions.

Literature and setting

The Dutch healthcare system (Kroneman et al., 2016) is characterized by mandatory basic health insurance. GPs serve as gatekeepers, meaning that specialist care is reimbursed only after a referral, and patients retain the freedom to choose their preferred hospital. Specialist care is delivered through various provider types: general hospitals, independent treatment centers (ITCs) offering one or more specialties, top-clinical hospitals for complex care, and teaching or academic hospitals for tertiary care. While hospitals offer a broad range of services, ITCs are often more specialized to perform mainly cataract or hip operations. Insurance companies contract specialist care providers, and care is reimbursed based on Diagnosis Treatment Combinations (DTCs). DTCs are similar to DRGs, contain information on the diagnosis, medical specialty, treatment, and price.

In 2018, Slotervaart Hospital, a general hospital in Amsterdam, The Netherlands, went bankrupt on October 25. The bankruptcy resulted from financial difficulties due to mismanagement, relatively low prices, and high fixed costs. As parts/blocks of the care were taken over by other hospitals, the Dutch Authority for Consumers and Markets (ACM) had to treat these reallocations of healthcare as a merger and conduct a merger investigation (ACM, 2019). The merger investigation was performed using patient flow analysis, requiring a geographic market definition and a product market definition (in this case groups of patients). As a result of this merger investigation, ACM permitted the specialist departments to be relocated to nearby hospitals. As a result of the bankruptcy, the existing and potential patients from Slotervaart hospital had to find an alternative hospital for their care. To mitigate endogeneity arising from patients' tendency to follow their specialist, we restrict the analysis to initial hospital choices—defined as new referrals from general practitioners, where patients have no prior relationship with a specialist—thereby minimizing this source of bias. (Beukers, Kemp, & Varkevisser, 2014). This allows us to exploit only the exogenous shock to patient preferences (one fewer alternative) and to compare the observed patient choices with the predicted patient choice.

Accurately predicting future patient flow is important for policymakers particularly in the context of evaluating hospital mergers (Capps, Dranove, & Satterthwaite, 2003) or health service planning

particularly when complex specialist care is centralized (Aggarwal et al., 2022). Mergers and decisions to centralize care significantly impact healthcare affordability as hospital mergers typically lead to higher prices (Gaynor, Ho, & Town, 2015) (Dafny, Ho, & Lee, 2019) (Brand, Garmon, & Rosenbaum, 2023) (Roos, Croes, Shestalova, Varkevisser, & Schut, 2019). In addition to price effects, mergers and centralization often negatively affect accessibility (Jiang, Fingar, Liang, Henke, & Gibson, 2021) (Aggarwal et al., 2022) and positive or negative implications for quality (Beaulieu et al., 2020) (Baum et al., 2022). Therefore, such decisions require thorough evaluations of the potential impact on these societal outcomes. During merger control proceedings, competition authorities such as the Federal Trade Commission (FTC) or the Dutch Authority for Consumers and Markets (ACM) assess the availability of competitors in the area, where patients can potentially seek care, if they are not satisfied with the merged hospital.

In order to assess the potential future effects of a merger, competition authorities often rely on diversion ratios. Diversion ratios are essentially a measure of substitutability from a patient perspective. The substitutability after the merger will decrease with the diversion ratio of the merging partner, allowing a merged hospital to increase its negotiated price; proportional to the diversion ratio between the merging hospitals (Garmon, 2017). High diversion ratios between the merging parties can therefore trigger merger control investigations (Garmon, 2017), underscoring their pivotal role in competition policy.

Over the past two to three decades, many countries have seen a large number of hospital mergers reviewed by competition authorities. There is extensive empirical literature showing that several approved mergers have led to higher prices, indicating that these transactions may have been anticompetitive. An illustrative example is the special issue of the International Journal of the Economics of Business (2011, vol. 18(1)), in which several FTC staff members present the results of event studies (Tenn, 2011) (Haas-Wilson & Garmon, 2011). (Elzinga & Swisher, 2011) argue that the widely used Elzinga-Hogaty test may not be appropriate for assessing hospital mergers. Studies outside the USA also showed that approved hospital mergers have led to relative price increases after completion (ACM, 2017) (Kemp, Kersten, & Severijnen, 2012) (Roos, Croes, Shestalova, Varkevisser, & Schut, 2019).

Moreover, the same methods used to obtain diversion ratios in merger control can inform decisions on health service planning, considering various outcomes such as price, quality or accessibility (Aggarwal et al., 2022). In terms of health service planning, when highly specialized care is concentrated, predicting future changes in patient flows is similarly important: Although quality may be higher in a farther away hospital, patients may not be able or willing to travel to these centers of excellence. Therefore, during the process of health service planning it's important to account for patient's preferences, when certain types of care are discontinued at various hospital locations. Many countries aim to centralize health services (Aggarwal et al., 2022) (Baum et al., 2022) (Versteeg, Ho, Siesling, & Varkevisser, 2018), largely based on medical literature showing a connection between volume and quality (Morche, Mathes, & Pieper, 2016). Demand models originally developed for merger control are increasingly being adapted for health service planning (Aggarwal et al., 2022), attempting to address the questions about the optimal tradeoff between distance and quality of care. Similarly to merger assessment methods, healthcare planning models also use choice modeling to

predict future demand, which can be used to construct diversion ratios that can inform the allocation of physicians or infrastructure. From a policy perspective, it is highly desirable that ex-ante diversion ratios closely match observed, ex-post patient flows. Accurate predictions enhance the ability of antitrust agencies and health policymakers to make informed decisions (Capps et al., 2003) (Balan & Brand, 2023) (Garmon, 2017). Ultimately, obtaining post-merger or post-centralization diversion ratios is a prediction problem (Athey & Imbens, 2019), comparing various models on their predictive performance and selecting the best model(s) may become feasible.

From a prediction perspective, hospital choice can be conceptualized as a multiclass classification problem, with specific hospitals being the different classes of the outcome variable. This classification problem can be solved using econometric or machine learning models. In multiple settings, machine learning models predict choice more accurately than traditional econometric models (Athey & Imbens, 2019) (Van Cranenburgh, Wang, Vij, Pereira, & Walker, 2021) (Zhao, Yan, Yu, & Van Hentenryck, 2020). To predict diversion ratios for hospitals within merger control, traditionally a patient flow analysis and increasingly choice models are used (Handel & Ho, 2021) (van der Geest & Varkevisser, 2024) (Raval et al., 2021) (Raval et al., 2022). In this study, we test whether merger control and health service planning can benefit from these machine learning models.

Methods

Data

We conducted this study on insurance claim data provided by Vektis business intelligence. Prediction models require a development set to build a prediction model and a test set (sometimes called a hold-out set) not used in the development of the prediction model to test the predictive performance of the prediction model (Hastie, Tibshirani, Friedman, & Friedman, 2009). In constructing the dataset, we followed the time frame used by competition authorities in merger assessments (U.S. Department of Justice and the Federal Trade Commission, 2023) when selecting the development and test set, using a relatively short period of time for both. The development set covered the period from January 2016 to July 2018, during which patient flows were stable prior to the bankruptcy of Slotervaart Hospital. The test set included data from January 2019 to December 2020, after the hospital had closed. To avoid potential endogeneity—where declining patient numbers might have contributed to the bankruptcy—we excluded the transition period between August and December 2018, during which patient volumes were already decreasing (Figure A1). This time frame also reflects a practical and ethical consideration to minimize the use of privacy-sensitive data while maintaining analytical rigor.

As the patient flow analysis requires a product market definition, we grouped services for groups of patients. The study population consisted of patients newly referred by general practitioners for cataract, ENT and intestinal cancer care, who had no previous relation with the specialist. These three patient groups were selected to reflect a range of clinical complexity and provider types. Cataract care is relatively low in complexity and often provided by both hospitals and independent treatment centers (ITCs), allowing us to examine how models differentiate between routine and more complex cases. ENT care spans a broader age range and clinical spectrum, providing a test of robustness across a heterogeneous population. Intestinal cancer care typically involves more specialized treatment, where

patients may be more willing to travel farther, making it a useful case for testing the influence of geographic variables. In addition, the choice set for intestinal cancer changed between the pre- and post-periods, offering a realistic scenario to assess model robustness.

To ensure meaningful market share estimates, we defined the relevant geographic market as the 80% service area of Slotervaart Hospital prior to its closure, including adjacent zip codes without patients to avoid blind spots. Providers with less than 1% market share in this area were excluded because such low-volume providers likely represent atypical cases (e.g., patients treated outside their usual region) and contribute minimally to overall diversion patterns. Hospitals that merged during the study period were treated as a single institution throughout. Each new referral from a GP was treated as a new hospital choice, and all analyses were conducted separately for each patient group, reflecting their distinct choice sets. The variables used in the models included hospital type, travel distance and time (measured from the centroid of the patient's four-digit postcode), gender, and age (calculated as the number of days between the treatment date and the patient's date of birth).

The claims data used in this study were provided by Vektis Business Intelligence. Due to legal restrictions, access to this analysis-dataset is restricted to ACM. However, researchers may construct the dataset and replicate the study using the exact same underlying microdata available through Statistics Netherlands (CBS), under secure access conditions.

Analyses

First, we used six methods to predict market share and the diversion ratios for the time-period following the bankruptcy.

1. Patient flow analysis based on market shares
2. Conditional logit
3. Mixed logit
4. LASSO
5. Random forest (RF)
6. Gradient boosting machines (GBM)

Our hypothesis was that all other methods would perform better than the patient flow analysis (Garmon, 2017).

In the patient flow analysis, the diversion ratio following the removal of Slotervaart Hospital is calculated using proportional allocation based on pre-closure market shares, as follows (Rossi, Whitehouse, & Moore, 2018) (Shapiro, 2010) (Willig, Salop, & Scherer, 1991):

$$Equation\ 1\ \widehat{D}^{jk} = \sum \frac{\hat{s}_{ik}}{1 - \hat{s}_{ij}}$$

Where $\widehat{D}^{jk}$ is the diversion ratio from j (Slotervaart) to hospital k, $\hat{s}_{ik}$ is the probability that patient *i* chooses hospital *k* and $\hat{s}_{ij}$ for hospital *j,* equal to the market share of the hospitals. In this case the market share $\hat{s}_{ik}$ for the future is assumed to be the same as before the merger. In other words, the market share after the merger is the same as the market share before the merger, and per patient the predicted probability of choosing a hospital does not change. The diversion ratio is commonly expressed as a percentage, and we follow that in this study. This is the simplest method to obtain diversion ratios, does not require patient-level information and will be used as a baseline for comparisons.

With the choice models, we predict the individual choice probabilities and aggregate them to obtain market shares for the development set. In the test set, in case the model predicted the Slotervaart hospital, we used the second choice to obtain market shares. The market shares yielded diversion ratios according to the formula above, using the predicted market shares. To fit the choice models, we used hospital-type dummies as alternative specific constants, age, sex, travel time by car and distance between the patient and hospital as patient-specific variables. We included both travel time and distance in order to account for the fact that in a large urban area cars may not be the preferred method of traveling to the hospital. Although including both travel time and distance introduces multicollinearity, it does not bias the coefficients and ensures consistency with machine learning models, which are not affected by this issue. In addition, we also included quadratic terms for time and distance in the choice models. We fitted the following conditional logit model with interactions between patient and hospital characteristics:

$$Equation\ 2\ U_{ij} = z_{ij}\alpha\ +\ x_i\delta_j\ + \gamma z_{ij}x_i\ + c_j\ + \epsilon_{ij}$$

where z_{ij} are hospital-specific variables and α coefficients for z, x_i patient-specific variables and δ_j coefficients for x and c_j the alternative specific constants and γ is the coefficient of $z_{ij}x_i$.and ϵ_{ij} the error terms.

In addition, we performed a mixed logit analysis assuming a correlation between time and distance:

$$Equation\ 3\ U_{ij} = w_{ij}\beta_i\ + z_{ij}\alpha\ +\ x_i\delta_j\ + \epsilon_{ij}$$

where w_{ij} are hospital-specific variables and β_i random coefficients that vary between patients.

Following the choice models, we also performed three machine learning models: LASSO, RFs, and GBM. A background on machine learning models is provided in the Appendix. Here, we repeatedly split the development set using a repeated cross-validation setup, where a random sample of 20% is used to split the development dataset into a training and a validation set, 5 times. The machine learning models use the validation sets to assess the accuracy of the models, and improve on them.

The first machine learning model was a LASSO regression applied to the conditional logit model (Friedman, Hastie, & Tibshirani, 2010). Briefly, the LASSO avoids overfitting and increases prediction accuracy by decreasing the size of small coefficients.

Subsequently, we fitted a RF (Breiman, 2001) (Hastie, Tibshirani, Friedman, & Friedman, 2009) using the previous variables. To maximize predictive performance on the test set, we used hyperparameter tuning, varying the number of trees between 100, 200, 500 and 1000, the number of variables between 5 and 6, and minimal node sizes (number of observations) between 2 and 10. The most accurate models were 500 trees, 6 variables, and a minimum node size of 2.

Finally, we fit a GBM model. GBM is similar to RF, only that it builds trees sequentially, considering the residuals in the previous step (Friedman, 2001). The following hyperparameters were used to obtain optimal predictive performance: 500 trees (varied between 100 and 500), interaction depth of 2 (varied between 1 and 3), shrinkage of 0.1 (varied between 0.1 and 0.2) and minimum number of observations (node sizes) between 2 and 10. In both models, we use grid search, which means that we tested all combinations of the hyperparameters specified previously (Hastie, Tibshirani, Friedman, & Friedman, 2009).

$$Equation\ 4\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

We treated all models as multiclass classifiers, assigning patients to hospitals based on predicted probabilities. Accuracy is defined as the proportion of correct predictions (True Positive and True negatives) out of the total number of predictions (true positives, true negatives, false positives and false negatives) made and is a key metric of machine learning models (Equation 4). We interpret the hit ratios as accuracy (Van Cranenburgh, Wang, Vij, Pereira, & Walker, 2021) and compute the coefficients and variable importance (Breiman, 2001), the counterpart of the coefficients in the RF and GBM models. Variable importance is a measure of the decrease of model performance (e.g. accuracy) if the variable is removed. Variable importance was assessed via permutation, measuring the drop in model accuracy when each variable was randomly shuffled (Breiman, 2001). The values are non-normalized raw effect sizes, calculated independently per variable and do not sum to 100%. Subsequently, we compared the models in terms of absolute difference between the predicted and observed diversion ratio. We evaluated model performance using two standard metrics: MAD and RMSE. MAD captures the average magnitude of prediction errors, while RMSE penalizes larger errors more heavily. These are defined as follows:

MAD = $1/n \sum|y_i - \hat{y}_i|\ from\ i = 1\ to\ n$

RMSE = $\sqrt{(1/n \sum(y_i - \hat{y}_i)^2\ from\ i = 1\ to\ n)}$

Where y_i is the observed diversion ratio and $\hat{y}_i$ is the predicted diversion ratio for hospital i.

We calculated the observed diversion ratio according to the formula used by Raval et.al. (Raval, Rosenbaum, & Wilson, 2021):

$$Equation\ 5\ Observed\ diversion\ ratio = \frac{s_k^{\text{post}} - s_k^{\text{pre}}}{s_j^{\text{pre}}}$$

where s_k^{post} is the post-bankruptcy share of hospital *k,* s_k^{pre} is the pre-bankruptcy share of hospital *k,* and s_j^{pre} is the pre-bankruptcy share of Slotervaart hospital *j*. By construction, the observed diversion ratio can be positive or negative, reflecting increases or decreases in patient share post-closure. We present the percentage point difference versus the observed diversion ratios by subtracting the observed diversion ratio from the diversion ratio obtained from the models.

Subsequently, we made a comparison in terms of RMSE, when compared to the patient flow analysis. Here, we use the following formula (Raval, Rosenbaum, & Wilson, 2021):

$$Equation\ 6\ \Delta RMSE = 1 - (\frac{RMSE_{s_{model}}}{RMSE_{s_{market\ share}}})$$

Where $RMSE_{s_{model}}$ is the RMSE of the conditional logit, mixed logit, RF and GBM methods, and $RMSE_{s_{market\ share}}$ is the RMSE of the patient flow analysis. By construction, values below 0 are worse than the patient flow analysis, 0 is the same as the patient flow analysis and 1 (100%) is perfect prediction.

In addition, we performed two sensitivity analyses: 1) an analysis where we extended the pre-period included in the development set (January 2014—July 1 2018) and the post-period in the test set (January 2019—December 2022) by two years, and 2) an analysis where we divided the dataset into nearby and faraway patients, based on the median driving time of 15 minutes.

All analyses were performed in R version 4.1.3.

Results

Descriptives

Our choice set included 30,316 cataract choices, 47,344 ENT choices, and 8,939 intestinal cancer choices in the pre-period, and 21,335, 34,278 and 5,374 choices respectively for the post period. Table A1 shows the number of claims per year. Hospital type and gender ratio were comparable between the pre- and post-periods across all patient groups. Travel time and distance slightly increased in the post-period, as did patient age (Tables 1 and 2).

[Table 1 about here]

[Table 2 about here]

Figure 1a and Figure 1b

Figure 1a displays provider types per choice set, providing insight into the competitive landscape. Figure 1b presents the number of patients per postcode area treated at Slotervaart Hospital during the pre-period, categorized by patient group. During the pre-period, Slotervaart Hospital held a market share of 5.7% for cataract referrals, 9.6% for ENT referrals, and 1% for intestinal cancer referrals. Table A2 in the appendix shows the market shares per hospital for all three groups of patients.

Main results of the development set

[Table 3 about here]

Table 3 shows the results of the conditional logit and mixed logit analyses for all three patient groups. Both the conditional logit and mixed logit models indicate a statistically significant negative relationship between distance and provider choice. For ENT and intestinal cancer patients, there is also a negative relationship between travel time and provider choice, while cataract patients appear willing to travel longer in terms of time, even if the physical distance is shorter. Furthermore, there is a significant relationship between hospital type and choice. The interaction between age and distance is significant for cataract and intestinal cancer, but not for ENT. The interaction between age and time is significant only in the mixed logit analysis of ENT. A significant relationship between gender and distance was observed only among cataract patients, suggesting that women are more willing to travel further for cataract surgery. Table A3 shows the shrunk coefficients for the LASSO algorithm.

[Table 4 about here]

Table 4 presents the relative variable importance scores for each predictor across the three patient groups—cataract, ENT, and intestinal cancer—using RF and GBM models. These scores reflect each variable's contribution to predictive performance and serve as a guide for variable inclusion. The values represent non-normalized raw effect sizes and do not sum to 100%. The most important variables are geographic variables. Distance consistently emerges as the most influential variable, followed by travel time, though the latter is notably less important in GBM models. Demographic variables such as age contribute less overall but still enhance predictive performance, particularly in the intestinal cancer group (2.7%). Institutional characteristics (type of hospital) also carry substantial predictive value and vary in importance across patient groups. The relevance of independent treatment centers (ITCs) is

especially pronounced for cataract patients, while academic hospitals appear most predictive for ENT care, as expected.

[Table 5 about here]

The accuracy of the different models is summarized above in table 5, for the development as well as for the test set. The accuracy scores are presented as percentages and represent the proportion of correctly predicted patient flows for each medical specialty within the development dataset, averaged across the 5 validation folds. In general, accuracy scores in the test set are not substantially lower than those in the development set, indicating that the model is not overfitted. For the RF and GBM models, the accuracy is similar for all three patient groups, with the RF model being somewhat better in the ENT group. On the contrary, the conditional logit and mixed logit models show lower accuracy scores, particularly in the cataract and intestinal cancer patient groups. Conditional logit performs slightly better for ENT patients and slightly worse for cataract than mixed models. The performance of the LASSO in the development set is slightly worse than the conditional logit. Accuracy of the models in the test set is similar or lower than in the development set, with the exception of the LASSO, where test set performance is higher.

Model performance

The MAD and RMSE values are calculated per hospital, and the figures summarize their distribution across the test set. This allows us to assess not only average performance but also variability in prediction accuracy.

[Figure 2 about here]

Figure 2 shows a comparison of the MADs between the predicted diversion ratios—based on the patient flow analysis, conditional logit, LASSO, mixed logit, RF, and GBM—and the observed diversion ratios for three distinct patient groups. Each boxplot displays the distribution of MADs across hospitals, with the thick line indicating the median error and the thin lines representing the interquartile range. Outliers are shown as individual points. The units are percentage points, and each point corresponds to a hospital in the test set. In all three patient groups, the RF method yielded the smallest average error, with GBM performing slightly worse. The other four methods—patient flow analysis, conditional logit, LASSO, and mixed logit—showed the largest errors in the cataract group. Specifically, mean MAD values for cataract were 48%–49% for the patient flow and choice models, compared to 3%–4% for the machine learning models.

[Figure 3 about here]

Figure 3 illustrates the relative improvement in RMSE across hospitals for each model compared to the baseline patient flow analysis. This combined Cleveland and box-and-whiskers plot displays the distribution of RMSE values in percentage points, where each point represents a hospital. The thick lines indicate the median improvement, thin lines show the interquartile range, and individual points denote outliers. Values above zero reflect better performance than the baseline. In terms of RMSE, the greatest improvement was achieved by the RF method, closely followed by GBM across all three patient groups. The choice models RMSE values were generally similar to the patient flow analysis, though it performed slightly worse for ENT patients. Notably, the machine learning models occasionally produced exact predictions. The largest difference in mean RMSE was observed in the Cataract group, with values ranging from 14%–17% for the patient flow and choice model, compared to just 1%–2% for the machine learning models.

Sensitivity analysis

Appendix 2 shows the effect of various alternative specifications on the main results. Figures A2 and A3 show the effects of including a longer pre- and post-period. The comparison between the prediction models is very similar to the base-case results. Compared to the base-case analysis, the predictive properties of the choice models deteriorate in all three patient groups, with the MAD values almost doubling. For the machine learning models, the deteriorations can also be seen, but the magnitude is somewhat smaller. When looking at RMSE values, these are slightly worse for the dataset containing a longer pre-post period, particularly for the cataract patient group.

Figures A4 and A5 in the appendix show the comparison between nearby patients and far-away patients. Compared to the base-case scenario, in all three patient groups we can see an increased MAD values, with the MAD values of the faraway patients being about 10-20 times the values of the nearby patients. In these datasets, the LASSO performs similarly to the other methods. RMSE is worse in both groups compared to the original dataset, in the nearby set it is twice as bad, in the faraway patient group the RMSE is 10-20 times as large. Comparing the model types, machine learning algorithms still perform better in the two groups than the market-share method, although this difference becomes marginal in the faraway group. The performance of the choice models is consistently worse than the patient flow analysis. The LASSO also performs differently compared to the base-case scenario, on average it is worse than the other models, with a far smaller interquartile range.

We performed another sensitivity analysis where we combined all three patient groups. Again, RF performed best, but the GBM, the two choice models and LASSO all performed worse than the patient flow analysis (Appendix figures A6-A7).

To support the main analyses, we also report MAD and RMSE results at the individual hospital level using Cleveland plots (Appendix Figures A8–A9), which confirm the relative performance patterns observed in the aggregate results.

Discussion

Summary of results

This study has compared six methods of calculating diversion ratios: 1) based on patient flow analysis, 2) conditional logit, 3) mixed logit, 4) LASSO, 5) RF and 6) GBM. These methods were compared on a dataset within a defined geographic, mostly urban area in and around Amsterdam. In addition, we compared the methods within specific patient groups. Of the methods considered, RF predicted the observed diversion ratio with the closest precision in all patient groups and this was consistent across all sensitivity analyses. In all patient groups the patient flow analysis performed the worst, with the exception of the sensitivity analysis combining all three patient groups.

Interpretation

We find that using RF leads to the smallest prediction error, thus outperforms the other methods. Extending the time frame reduced model accuracy, likely due to changes in the choice set that introduced additional variability, supporting the use of a relatively short timeframe for the analysis. The comparability of choice models and patient flow analysis is consistent with (Rossi, Whitehouse, & Moore, 2018), who also found that GP referral analysis based on patient flow analysis were similar to choice models (Garmon, 2017).

Our study is also similar to Raval (Raval, Rosenbaum, & Wilson, 2021), who compared machine learning models and choice models in the context of a disaster scenario. Raval et al. finds that machine learning models perform somewhat better when the choice set has small changes, and choice models perform better when the choice set has large changes (50-80% of the choice set is destroyed). Our results align more with the scenario investigated by Raval et al. when the choice set has small changes. To further explore the limits of machine learning models, it is instructive to compare the setting between a disaster-scenario and a case such as ours, which can be realistically expected when performing a merger assessment or a health service planning exercise. There are three aspects which impact the performance of machine learning models: 1) stability of the choice set, 2) stability of patient population in terms of travel time and 3) how the dataset was constructed.

In contrast to the Sumter Regional Hospital tornado scenario (Americus, GA, March 2007), which caused the largest change in the choice set (destroyed hospital share ~50%) where for some patients 50-80% of the choice set is removed, the choice set in this current study was largely unchanged. In addition to the change in the set of choices caused by the Slotervaart bankruptcy, there was no change in cataract and ENT providers and there was only a limited change in intestinal cancer providers. Such a pattern in the data is consistent with merger control investigations, where the expected changes are limited. When considering healthcare planning scenarios, major changes in the choice set are only possible in extreme cases when certain healthcare services which were previously available in a large number of hospitals are concentrated to a handful centers of excellence.

Second, our results are consistent with the literature finding that travel time/distance is the most important predictor in choice models (Capps, Dranove, & Satterthwaite, 2003) (Versteeg, Ho, Siesling, & Varkevisser, 2018) (Aggarwal et al., 2022), but it is unclear if it aligns with the results of the Raval study, where the variable importance of distance is not reported (Raval, Rosenbaum, & Wilson, 2021). In the disaster context, the patients themselves are also frequently forced to move, which may be a factor in the performance of the machine learning models after a hurricane or earthquake. If the patients were no longer at their home address, the nearest hospital will change, which is a change in the data that machine learning models cannot account for. This is important, as administrative data do not typically contain temporary addresses of the people and uses inaccurate distance data. While (Raval, Rosenbaum, & Wilson, 2021) perform a robustness check between areas impacted by a disaster (where patients were forced to move) and areas not impacted, it does not account for which patients actually moved, which may have happened as a precautionary measure in areas not directly impacted by a disaster as well. Furthermore, the removal of impacted areas not only shows a difference in the model performance, but also introduces additional complications, such as changes in patient flow analysis (Raval, Rosenbaum, & Wilson, 2021). In contrast, in our current study, the most important predictors (distance and time) did not change between the pre- and post periods, allowing the machine learning models to perform better than the choice models.

Third, the results depend on the underlying data. Compared to Raval (Raval, Rosenbaum, & Wilson, 2021) , who compared the models on a binned dataset, grouping together patients with similar age and distance categories, combined with hospital type and other characteristics, we focused on creating a unique observation per patient by increasing the precision of the travel time, distance and age variables, which were exactly computed. The granular data, combined with the non-linear relationships allowed the RF and GBM models to reach less biased predictions. In addition, this study, the only time when RF predictions were not clearly superior to the patient flow analysis is the case of the faraway patients. There, all models had major difficulty in predicting the diversion ratios. A possible explanation for that could be that those patients live outside of the urban core of Amsterdam, and subjective terms such as “near” and “far” may lead to other choice of providers. This dense urban setting may explain the better performance of machine learning models, as the granularity may be necessary to capture the complexity of these settings, in contrast to more general settings of Raval et.al. (Raval, Rosenbaum, & Wilson, 2021) where choice models perform adequately. Another factor relates to the dataset’s spatial structure: it resembles a ring of postcodes located 15–30 minutes from Slotervaart Hospital, with no observations from the immediate surrounding area.

Policy implications and extensions

While choice models are grounded in economic theory, machine learning models prioritize predictive accuracy, which may offer practical advantages in policy applications. There were numerous critiques over the years in the choice-modelling literature (Van Cranenburgh, Wang, Vij, Pereira, & Walker, 2021) concerning prediction models. In the past explainability was lacking, making acceptance of these models by researchers and policy-makers, despite potential superior predictive performance, more difficult. However, this has been remedied with variable importance measures (Van Cranenburgh, Wang, Vij, Pereira, & Walker, 2021), which we have also used in this paper. In addition, machine

learning models allow us to exploit a possibly nonlinear relationship between travel time and physical distance, which seems to drive the better predictive performance of the RF and GBM models.

An important consideration in regulatory decision-making is the asymmetry in the consequences of prediction errors. As event studies on hospital mergers show, approving a merger that should have been blocked may lead to irreversible harm, such as higher prices (Tenn, 2011) (Haas-Wilson & Garmon, 2011) (Elzinga & Swisher, 2011)(ACM, 2017) (Kemp, Kersten, & Severijnen, 2012) (Roos, Croes, Shestalova, Varkevisser, & Schut, 2019) (Brand, Garmon, & Rosenbaum, 2023) reduced accessibility (Jiang, Fingar, Liang, Henke, & Gibson, 2021) (Aggarwal et al., 2022),or lower quality of care (Beaulieu et al., 2020) (Baum et al., 2022).

. And a consummated hospital merger is difficult to reverse. While not conclusive, this general trend suggests that the use of less accurate models could make it more difficult for enforcement bodies to make correct enforcement decisions(Raval, Rosenbaum, & Wilson, 2016). In contrast, blocking a merger that would have generated efficiency gains may delay or forgo potential benefits, but these effects are generally less harmful to consumers in the short term. Therefore, minimizing false positives (i.e., erroneous approvals) is particularly important in this context. Our findings suggest that machine learning models, particularly RF, reduce prediction error and may thus help regulators avoid such costly mistakes.

Our method aligns with standard approaches used in merger evaluations and hospital centralization studies. Therefore, for hospital mergers and studies on hospital centralization, we would recommend using RF.

Our results can be extended for other policy-relevant applications within merger control. More reliable market share predictions have other uses beyond diversions, such as using Willingness to pay WTP analysis to evaluate hospital mergers (Garmon, 2017). Additionally, incorporating more precise diversion ratios into WTP and Upward Pricing Pressure UPP models allows for more sophisticated merger simulations (Balan & Brand, 2023), which may potentially predict prices after mergers more accurately (Garmon, 2017).

Strengths and limitations

A major strength of this study is that it models a realistic case when the choice set changes due to bankruptcy, an event similar to the situation when a merged provider closes a hospital location. Also, in contrast to a more data-driven machine learning studies, in this current study, we included variables determined by the choice models, which was more hypothesis-driven and supported by domain knowledge. As a result, the machine learning models drew on variables which were shown to consistently predict hospital choice in many studies, and potentially amplified the predictive power. Such an analysis can be performed on claims data, or more widely available discharge data containing patient address and various patient characteristics, as well as hospital characteristics. When providers discontinue a service, centralize specialist care, but also in merger cases it is a distinct possibility that the care at the old hospital location is not available after the event. Furthermore, the results held even considering the major disruptions in patient flows due to COVID (De Graaff et al., 2022). Furthermore,

in the intestinal cancer setting, three other alternative providers exited the market in the post period, so we could also exploit the effect of a somewhat dynamic choice set, further supporting the robustness of the results. Currently, machine learning models can be used to predict the market structure after a merger or centralization, as well as the diversion ratios. On a practical level, machine learning models do not require the construction of a choice set with all possible alternatives, allowing the analysis to be performed on a smaller dataset, potentially reducing analysis time.

The first limitation of this study is that it only looked into one hospital bankruptcy, while hospital closure may be more prominent. Bankruptcies are relatively rare in the Dutch healthcare sector, and bankrupt hospitals are usually merged with nearby hospitals, keeping their original location open. The Slotervaart bankruptcy is the only case when an entire hospital completely disappeared from its original location. Using this case allows us to test the model on an exogenous case, in contrast to more gradual process that may lead to hospital closure, where the effects of the closure itself cannot be identified due to the large changes in choice set. Therefore, the exact results may be different in other hospitals. In addition, it only examined three groups of patients, and in other groups of patients different results may emerge. Second, predictions suffer in quality when looking only at patients who live far away from Slotervaart. Further research is needed to uncover the reasons for this pattern. Another limitation of this study is that it does not include an outside option. Establishing travel time and travel distance for an outside option would be highly inaccurate, and in an urban area such as Amsterdam, only a handful of patients make use of the outside option, primarily when needing medical care on holiday or when visiting family. In addition, the inelasticity of healthcare (Koc, 2004) leads to the fact that patients in general will not forgo healthcare if the most preferred alternative is no longer available.

Finally, while our findings are robust within the context of the Slotervaart Hospital closure, their generalizability to other hospital markets or time periods may be limited. The Dutch healthcare system is characterized by strong GP gatekeeping, dense urban hospital networks, and relatively short travel distances, which may not reflect conditions in more rural or decentralized systems. Moreover, the Slotervaart case involved a complete and abrupt hospital closure, whereas many mergers or centralization efforts retain physical locations and staff, potentially leading to different patient responses. As such, while the predictive superiority of machine learning models is evident in this setting, further validation in other institutional and geographic contexts is warranted.

Future research

This study could be extended to include other patient groups as well. Furthermore, in addition to merger cases, predicting patient flows is also interesting in other settings, such as when health systems concentrate highly specialized care and hospitals close specific departments at certain locations. Therefore, future research should explicitly focus on how well choice models or machine learning models predict patient flows after centralizing care. In addition to these machine learning models, future research using more complex deep-learning models may be beneficial. This will especially be useful if such models gain on interpretability in the future.

Conclusion

In essence, this paper provides a framework for selecting and validating prediction models to predict future diversion ratios within the context of hospital merger evaluation and health services planning. Specifically, in the case of the Slotervaart bankruptcy, RF has performed slightly better than GBM and far better than conditional logit, mixed logit, and patient flow analysis in predicting observed diversion ratios. Although the theoretical advantages of choice modelling are clear, the predictive accuracy of the RF model is far superior. In addition, the market shares required for diversion ratios and various other merger control indicators, such as the WTP estimates or UPP estimates are also yielded by this model. When the goal is to predict future market shares and diversion ratios, RF appears to be the most effective method.

References

ACM. (2017). *Price and volume effects of hospital mergers.* The Hague, The Netherlands: ACM

ACM. (2019). *Goedkeuring van de concentratie tussen stichting OLVG en afdelingen MC Slotervaart .* The Hague, The Netherlands: ACM.

Aggarwal, A., Han, L., Van Der Geest, S., Lewis, D., Lievens, Y., Borras, J., . . . Van Der Meulen, J. (2022). Health service planning to assess the expected impact of centralising specialist cancer services on travel times, equity, and outcomes: A national population-based modelling study. *The Lancet Oncology, 23*(9), 1211–1220.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics, 11*, 685–725.

Balan, D. J., & Brand, K. (2023). Simulating hospital merger simulations. *The Journal of Industrial Economics, 71*(1), 47–123.

Baum, P., Lenzi, J., Diers, J., Rust, C., Eichhorn, M. E., Taber, S., . . . Wiegering, A. (2022). Risk-adjusted mortality rates as a quality proxy outperform volume in surgical oncology—A new perspective on hospital centralization using national population-based data. *Journal of Clinical Oncology, 40*(10), 1041–1050.

Beaulieu, N. D., Dafny, L. S., Landon, B. E., Dalton, J. B., Kuye, I., & McWilliams, J. M. (2020). Changes in quality of care after hospital mergers and acquisitions. *New England Journal of Medicine, 382*(1), 51–59.

Beukers, P. D., Kemp, R. G., & Varkevisser, M. (2014). Patient hospital choice for hip replacement: Empirical evidence from the netherlands. *The European Journal of Health Economics, 15*, 927–936.

Brand, K., Garmon, C., & Rosenbaum, T. (2023). In the shadow of antitrust enforcement: Price effects of hospital mergers from 2009 to 2016. *The Journal of Law and Economics, 66*(4), 639–669.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Capps, C., Dranove, D., & Satterthwaite, M. (2003). Competition and market power in option demand markets. *RAND Journal of Economics, *, 737–763.

Dafny, L., Ho, K., & Lee, R. S. (2019). The price effects of cross-market mergers: Theory and evidence from the hospital industry. *The Rand Journal of Economics, 50*(2), 286–325.

De Graaff, M. R., Hogenbirk, R. N., Janssen, Y. F., Elfrink, A. K., Liem, R. S., Nienhuijs, S. W., . . . Melenhorst, J. (2022). Impact of the COVID-19 pandemic on surgical care in the netherlands. *British Journal of Surgery, 109*(12), 1282–1292.

Elzinga, K. G., & Swisher, A. W. (2011). Limits of the Elzinga–Hogarty test in hospital mergers: The evanston case. *International Journal of the Economics of Business, 18*(1), 133–146.

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

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*, 1–22.

Garmon, C. (2017). The accuracy of hospital merger screening methods. *The Rand Journal of Economics, 48*(4), 1068–1102.

Gaynor, M., Ho, K., & Town, R. J. (2015). The industrial organization of health-care markets. *Journal of Economic Literature, 53*(2), 235–284.

Haas-Wilson, D., & Garmon, C. (2011). Hospital mergers and competitive effects: Two retrospective analyses. *International Journal of the Economics of Business, 18*(1), 17–32.

Handel, B., & Ho, K. (2021). The industrial organization of health care markets. *Handbook of industrial organization* (pp. 521–614) Elsevier.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* Springer.

Heida, J. P., van Engelsen, B., Baeten, S., & van Gent, C. (2016). *Product market definition in hospital care.* The Hague, The Netherlands: SIRM. Verkregen van https://www.acm.nl/sites/default/files/documents/2018-12/product-market-definition-in-hospital-care-in-the-netherlands.pdf

Jiang, H. J., Fingar, K. R., Liang, L., Henke, R. M., & Gibson, T. P. (2021). Quality of care before and after mergers and acquisitions of rural hospitals. *JAMA Network Open, 4*(9), e2124662.

Kemp, R. G., Kersten, N., & Severijnen, A. M. (2012). Price effects of dutch hospital mergers: An ex-post assessment of hip surgery. *De Economist, 160*(3), 237–255.

Koc, C. (2004). A theoretical rationale for an inelastic demand for health care. *Economics Letters, 82*(1), 9–14.

Kroneman, M., Boerma, W., van den Berg, M., Groenewegen, P., de Jong, J., van Ginneken, E., & World Health Organization. (2016). Netherlands: Health system review.

Morche, J., Mathes, T., & Pieper, D. (2016). Relationship between surgeon volume and outcomes: A systematic review of systematic reviews. *Systematic Reviews, 5*, 1–15.

Raval, D., Rosenbaum, T., & Wilson, N. (2016). Industrial reorganization: Learning about patient substitution patterns from natural experiments.

Raval, D., Rosenbaum, T., & Wilson, N. E. (2022). Using disaster-induced closures to evaluate discrete choice models of hospital demand. *The Rand Journal of Economics, 53*(3), 561–589.

Raval, D., Rosenbaum, T., & Wilson, N. E. (2021). How do machine learning algorithms perform in predicting hospital choices? evidence from changing environments. *Journal of Health Economics, 78*, 102481. doi:10.1016/j.jhealeco.2021.102481

Roos, A., Croes, R. R., Shestalova, V., Varkevisser, M., & Schut, F. T. (2019). Price effects of a hospital merger: Heterogeneity across health insurers, hospital products, and hospital locations. *Health Economics, 28*(9), 1130–1145.

Rossi, C., Whitehouse, R., & Moore, A. (2018). *Estimating diversion ratios in hospital mergers.* London, UK: CMA. Verkregen van https://assets.publishing.service.gov.uk/media/5f06e22fd3bf7f2bef137641/Estimating_diversion_ratios_in_hospital_mergers.pdf

Shapiro, C. (2010). The 2010 horizontal merger guidelines: From hedgehog to fox in forty years. *Antitrust LJ, 77*, 49.

Tenn, S. (2011). The price effects of hospital mergers: A case study of the Sutter–Summit transaction. *International Journal of the Economics of Business, 18*(1), 65–82.

U.S. Department of Justice and the Federal Trade Commission. (2023). *Merger guidelines*. Washington D.C.: DOJ & FTC. Verkregen van https://www.ftc.gov/system/files/ftc_gov/pdf/2023_merger_guidelines_final_12.18.2023.pdf

Van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., & Walker, J. (2021). Choice modelling in the age of machine learning. *arXiv Preprint arXiv:2101.11948,*

van der Geest, S. A., & Varkevisser, M. (2024). Steering them softly with a quality label? A case study analysis of a patient channelling strategy without financial incentives. *The International Journal of Health Planning and Management,*

Versteeg, S. E., Ho, V., Siesling, S., & Varkevisser, M. (2018). Centralisation of cancer surgery and the impact on patients' travel burden. *Health Policy, 122*(9), 1028–1034.

Willig, R. D., Salop, S. C., & Scherer, F. M. (1991). Merger analysis, industrial organization theory, and merger guidelines. *Brookings Papers on Economic Activity.Microeconomics, 1991*, 281–332.

Zhao, X., Yan, X., Yu, A., & Van Hentenryck, P. (2020). Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel Behaviour and Society, 20*, 22–35.

Tables

Table 1 Hospital level characteristics

	Pre (January 1 2016-31 June 2018			Post (January 1 2019-December 31 2020)		
Cataract						
	Number of claims	Percent claims	Size of choice set per category	Number of claims	Percent claims	Size of choice set per category
Top-clinical	1111	3.66	1	845	3.96	1
Teaching hospital	8378	27.64	1	4808	22.54	1
ITC	15589	51.42	5	13668	64.06	5
General hospital	5238	17.28	3	2014	9.44	2
ENT						
	Number of claims	Percent claims	Size of choice set per category	Number of claims	Percent claims	Size of choice set per category
Teaching hospital	14420	30.46	1	10997	32.08	1
Academic	3906	8.25	2	2166	6.32	2
ITC	7965	16.82	3	8048	23.48	3
General hospital	21053	44.47	5	13067	38.12	4
Intestinal cancer						
	Number of claims	Percent claims	Size of choice set per category	Number of claims	Percent claims	Size of choice set per category
Top-clinical	659	7.37	1	638	11.87	1
Teaching hospital	2017	22.56	1	1450	26.98	1
Academic	412	4.61	2	280	5.21	2
ITC	2013	22.52	5	1376	25.6	3
General hospital	3622	40.52	6	1404	26.13	4
Other	216	2.42	1	226	4.21	1

ENT= ear, nose and throat, ITC=Independent treatment center

Table 2 Patient level characteristics

	Pre						Post						
Cataract													
	Number	Perc.	Mean	SD	Min	Max	Number	Perc.	Mean	SD	Min	Max	p-values (t-test)
Patients	30316						21335						
Female	17779	59%					12311	58%					
Distance (km)			15.31	16.09	0.00	65.61			18.33	16.85	0.00	67.07	<0.001
Time (min driving)			16.26	11.61	0.00	48			18.48	12.17	0.00	47	<0.001
Age			72.09	10.26	0.13	102.86			72.29	10.05	0.16	104.03	0.004
ENT													
	Number	Perc.	Mean	SD	Min	Max	Number	Perc.	Mean	SD	Min	Max	
Patients	47344						34278						
Female	23773	50%					17189	50%					
Distance (km)			8.05	7.81	0	57.83			9.35	8.86	0	61.15	<0.001
Time (min driving)			11.02	6.75	0	45			12.11	7.27	0	49	<0.001
Age			32.76	25.52	0.04	105.14			34.95	24.83	0.05	103.00	<0.001
Intestinal cancer													
	Number	Perc.	Mean	SD	Min	Max	Number	Perc.	Mean	SD	Min	Max	
Patients	8939						5374						
Female	4171	47%					2544	47%					
Distance (km)			11.19	10.06	0	82.36			13.30	10.53	0	70.00	<0.001
Time (min driving)			13.21	7.45	0	55			14.62	7.57	0	46	<0.001
Age			63.38	12.50	0.59	100.59			63.80	12.52	1.00	100.81	0.031

Table 3 Results of the conditional logit and mixed logit analyses in the development set

	Cataract		ENT		Intestinal Cancer	
	Conditional Logit	**Mixed Logit**	**Conditional Logit**	**Mixed Logit**	**Conditional Logit**	**Mixed Logit**
Variable						
Distance (km)						
Linear Term	-0.353*** (0.027)	-0.352*** (0.027)	-0.178*** (0.008)	-0.087*** (0.010)	-0.124*** (0.030)	-0.178*** (0.049)
Quadratic Term	0.001*** (0.0001)	0.001*** (0.0001)	-0.001*** (0.0001)	-0.004*** (0.0002)	-0.0004** (0.0002)	-0.010*** (0.001)
Academic*Distance	0.168*** (0.012)	0.167*** (0.011)	0.135*** (0.010)	0.104*** (0.009)	0.191*** (0.019)	0.130*** (0.020)
Top-clinical*Distance	0.110*** (0.020)	0.109*** (0.020)			0.050** (0.025)	0.008 (0.031)
Teaching*Distance	0.168*** (0.012)	0.167*** (0.011)	0.108*** (0.006)	0.069*** (0.007)	0.111*** (0.012)	0.089*** (0.014)
ITC*Distance	0.241*** (0.012)	0.241*** (0.012)	0.178*** (0.008)	0.104*** (0.009)	0.168*** (0.013)	0.111*** (0.014)
Other*Distance					0.128*** (0.023)	0.086*** (0.023)
Age*Distance	-0.001*** (0.0004)	-0.001*** (0.0004)	0.0002* (0.0001)	-0.0001 (0.0001)	-0.001*** (0.0004)	-0.003*** (0.001)
Female*Distance	0.018** (0.007)	0.018** (0.007)	-0.001 (0.006)	0.001 (0.007)	0.011 (0.011)	-0.0005 (0.017)
Time						
Linear Term	0.069** (0.033)	0.068** (0.033)	-0.139*** (0.010)	-0.233*** (0.012)	-0.045 (0.038)	-0.124** (0.055)
Quadratic Term	-0.001*** (0.0002)	-0.001*** (0.0002)	0.0002 (0.0002)	-0.0003 (0.0003)	-0.001 (0.0004)	0.002*** (0.001)
Academic*Time			-0.075*** (0.011)	-0.013 (0.011)	-0.074*** (0.024)	0.015 (0.026)
Top-clinical*Time	0.079*** (0.019)	0.081*** (0.020)			0.020 (0.028)	-0.051 (0.039)
Teaching*Time	-0.033*** (0.013)	-0.032*** (0.012)	0.063*** (0.007)	0.122*** (0.008)	0.069*** (0.014)	0.082*** (0.019)
ITC*Time	-0.072*** (0.012)	-0.072*** (0.012)	-0.068*** (0.009)	-0.006 (0.010)	0.006 (0.016)	0.057*** (0.019)
Other*Time					0.074** (0.030)	0.117*** (0.030)
Age*Time	-0.0003 (0.0004)	-0.0003 (0.0004)	-0.0001 (0.0001)	0.0003** (0.0002)	0.0002 (0.001)	-0.0004 (0.001)
Female*Time	-0.027*** (0.009)	-0.027*** (0.009)	0.004 (0.007)	0.002 (0.009)	-0.020 (0.014)	-0.031* (0.018)
Interaction Terms						
Distance*Distance		-0.024 (0.018)		0.102*** (0.008)		0.380*** (0.023)
Distance*Time		0.042* (0.024)		0.021** (0.010)		-0.107*** (0.021)
Time*Time		0.002 (0.006)		0.057*** (0.006)		0.114*** (0.015)
Observations	30,767	30,767	47,344	47,344	8,982	8,982
R2	0.421	0.421	0.504	0.510	0.328	0.340
Log Likelihood	-37,753.580	-37,752.230	-50,446.430	-49,810.320	-15,082.250	-14,815.060
LR Test	54,807.070*** (df = 41)	54,809.770*** (df = 44)	102,470.700*** (df = 44)	103,742.900*** (df = 47)	14,728.530*** (df = 63)	15,262.900*** (df = 66)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4 Variable importance* for the RF and GBM models

	Cataract		ENT		Intestinal cancer	
	RF tuned	GBM	RF tuned	GBM	RF tuned	GBM
Age	0.285%	0.743%	0.355%	0.271%	0.608%	2.700%
Female	0.038%	0.013%	0.031%	0.003%	0.087%	0.143%
Time	34.297%	12.319%	33.263%	5.310%	26.910%	8.536%
Distance (km)	46.723%	43.996%	42.923%	38.723%	39.081%	36.815%
Top-clinical	4.267%	6.045%			11.130%	9.631%
Teaching	37.149%	30.575%	40.060%	32.900%	31.250%	23.924%
Academic			10.910%	9.415%	5.113%	5.086%
ITC	27.953%	6.308%	22.081%	13.469%	24.835%	9.769%
Other					3.180%	3.336%

*decrease in predictive accuracy if the variable is removed.

Table 5 Accuracy of the models on the development and test sets

Development set				Test set		
	Cataract	ENT	Intestinal cancer	Cataract	ENT	Intestinal cancer
RF	0.96	0.96	0.93	0.96	0.95	0.94
GBM	0.96	0.94	0.93	0.96	0.94	0.94
Conditional logit	0.51	0.64	0.44	0.44	0.64	0.45
LASSO	0.45	0.59	0.42	0.49	0.64	0.47
Mixed logit	0.54	0.63	0.44	0.46	0.65	0.45

Figures

Figure 1 a Location of the hospitals within the geographic market

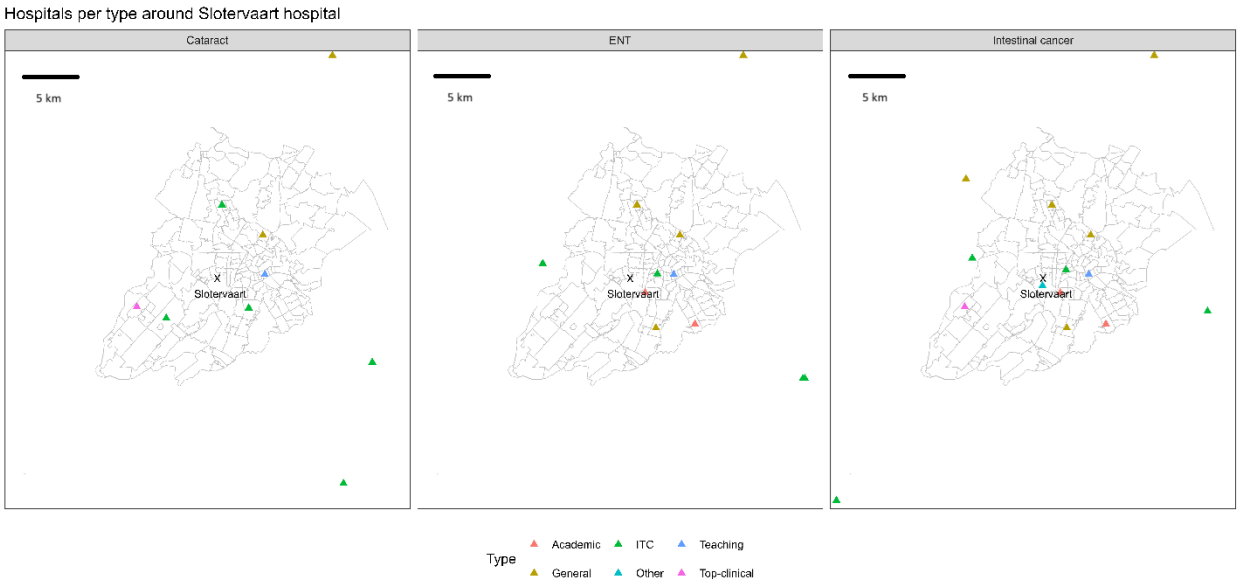


Figure 1 b Number of patients per 4 level postcode

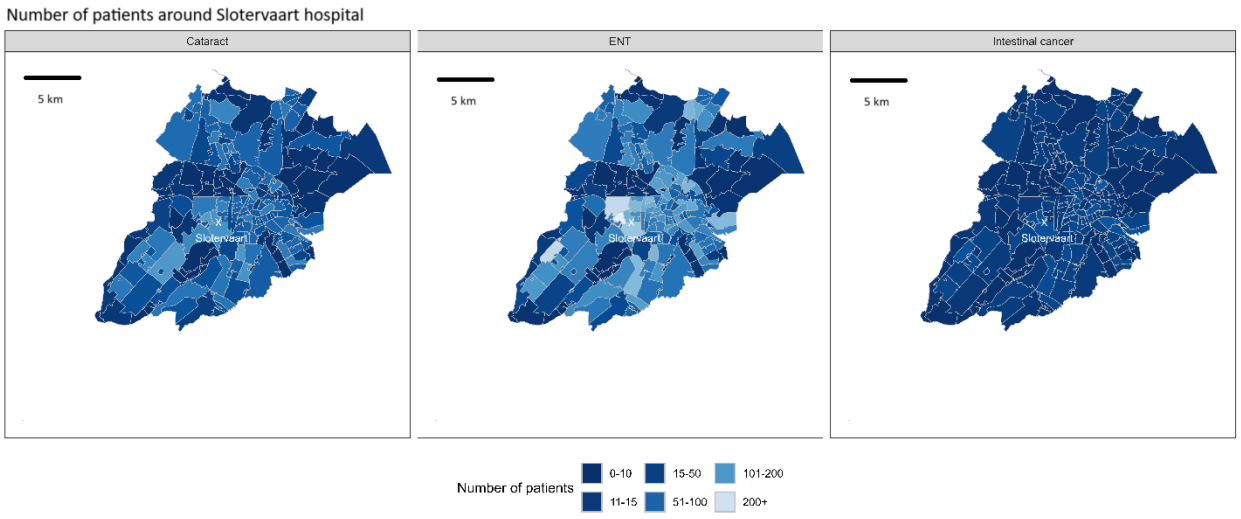
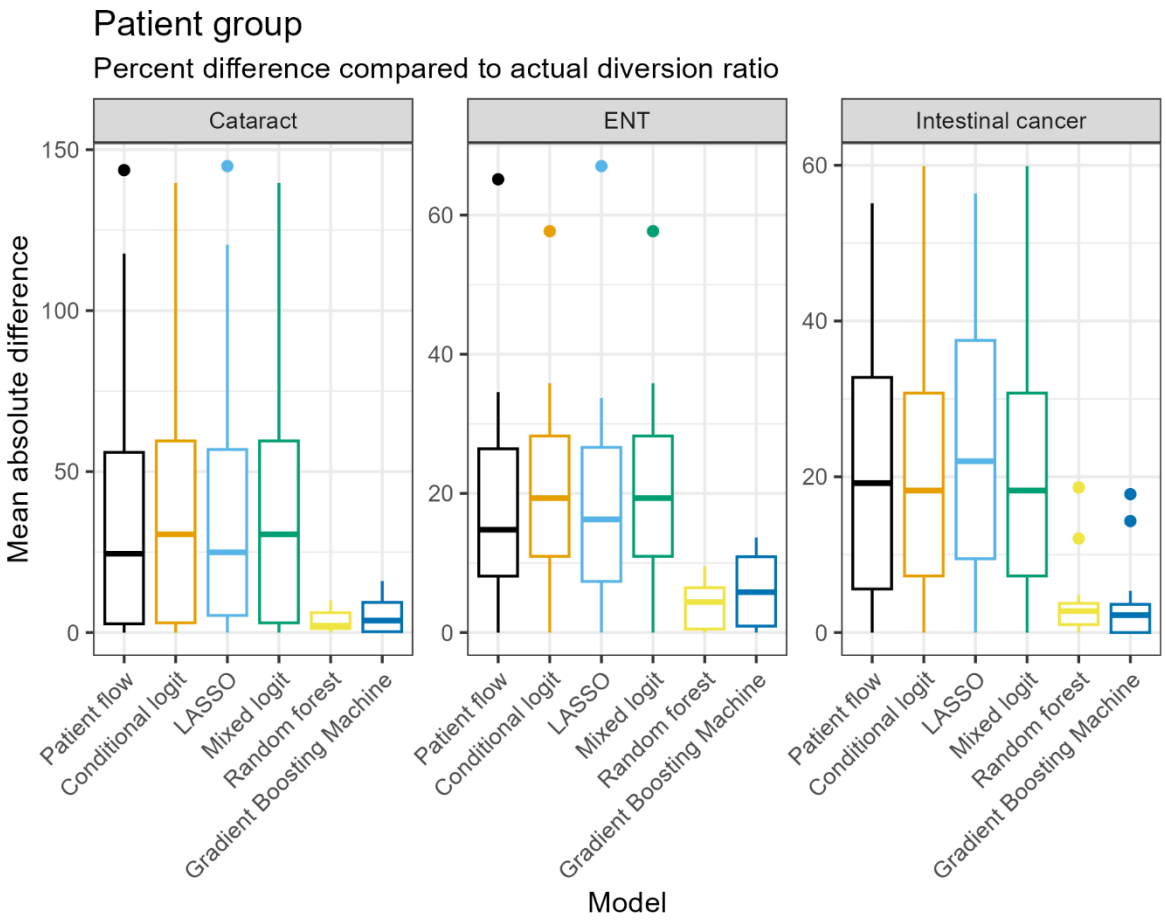
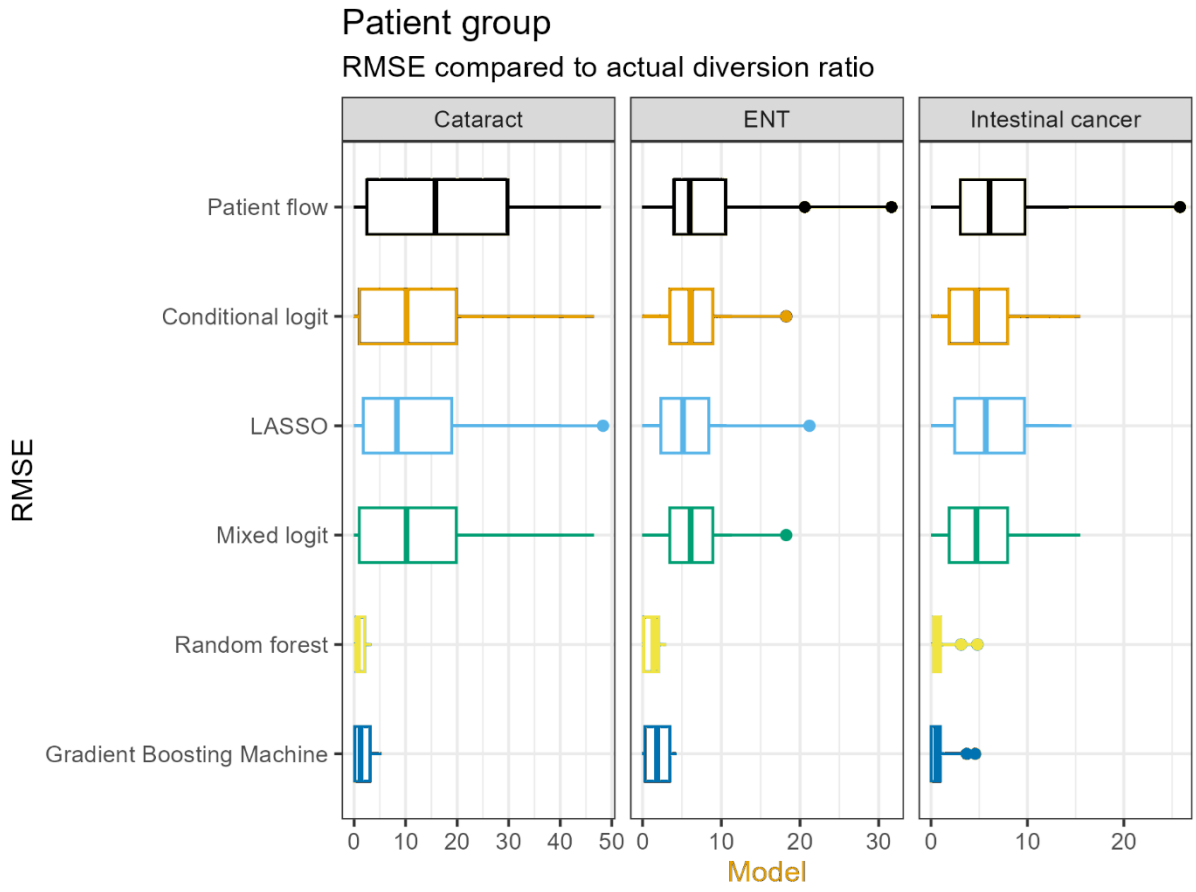


Figure 2 Mean Absolute difference between estimated and actual diversion ratios for three patient groups



Distribution of absolute errors between predicted and observed diversion ratios across hospitals (n = 20). Boxplots show median, interquartile range, and outliers. Units are percentage points.

Figure 3. Percent improvement in RMSE versus the observed diversion ratios

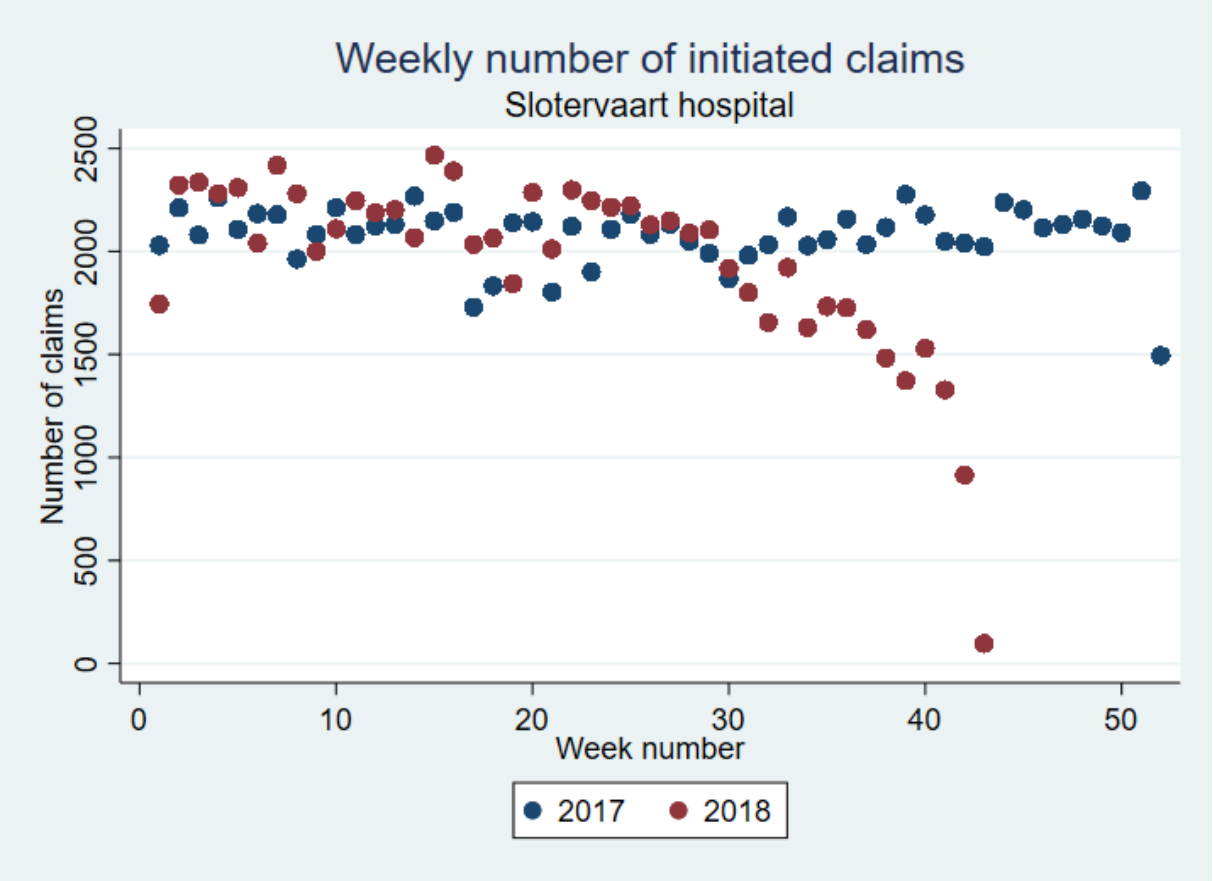


Distribution of RMSE values for each model across hospitals. RMSE is calculated per hospital and expressed in percentage points. Models are color-coded consistently with Figure 2.

Appendix

Appendix 1

Figure A1 Weekly number of claims



Appendix 2 Background on Prediction models, LASSO, RF and GBM models.

1. Basic set-up of the prediction dataset

The initial dataset is split into a development set, where the model will be developed, and a test set, or hold-out set for final evaluation. The development set is further divided into a training and validation sets.

Within the development set, in order to assess the performance of the model, 5-fold cross-validation is employed. This technique involves dividing the development set into five equally sized subsets or "folds." The model is trained and validated five times, each time using a different fold as the validation set and the remaining four folds as the training set. The performance metrics, such as accuracy, precision, and recall, are averaged across the five iterations to obtain a more robust estimate of the model's performance. This approach is well-documented in the literature, with Athey & Imbens (2019) highlighting its effectiveness in ensuring model generalization. Kohavi (1995) also emphasizes the importance of cross-validation in providing a reliable estimate of model performance, while Refaeilzadeh et al. (2009) discuss its role in preventing overfitting.

During each fold of the cross-validation, the model is trained on the training set, and its performance is evaluated on the validation set. Hyperparameters are tuned based on the validation performance to optimize the model's configuration. This process is crucial for preventing overfitting and ensuring that the model performs well on unseen data.

Once the model has been trained and validated using cross-validation, it is evaluated on the test set. The test set provides an unbiased estimate of the model's performance on new, unseen data, allowing for a final assessment of its effectiveness.

Overall, this setup ensures that the model is trained effectively, hyperparameters are tuned properly, and the model's performance is evaluated accurately, aligning with best practices in machine learning.

2. LASSO

LASSO, which stands for Least Absolute Shrinkage and Selection Operator, is a type of linear regression that incorporates regularization to enhance the prediction accuracy and interpretability of the statistical model it produces. It is particularly useful when dealing with datasets that have a high number of predictors.

LASSO extends traditional linear regression by adding a regularization term to the loss function used in ordinary least squares regression. This regularization term is known as the L1 penalty, which imposes a cost on the absolute size of the regression coefficients. The amount of regularization applied is controlled by a tuning parameter, often denoted by lambda (Friedman, Hastie, & Tibshirani, 2010).

The L1 penalty has two main effects. First, it shrinks the coefficients of less important features towards zero, which can reduce the variance of the estimates and potentially improve prediction accuracy. Second, it can shrink some coefficients to exactly zero, effectively performing feature selection by excluding irrelevant variables from the model. This is particularly useful in high-dimensional settings where the number of predictors exceeds the number of observations (Hastie, Tibshirani, & Friedman, 2009).

The optimization problem for LASSO is typically solved using coordinate descent, an iterative algorithm that updates one coefficient at a time while holding the others fixed. This approach is computationally efficient and well-suited for high-dimensional data (Friedman, Hastie, & Tibshirani, 2010).

The resulting LASSO model is often easier to interpret than traditional linear regression models because it tends to produce sparser models with fewer non-zero coefficients. This can make it easier to identify the most important predictors (Hastie, Tibshirani, & Friedman, 2009).

In summary, LASSO extends traditional linear regression by incorporating regularization through the L1 penalty, which helps to improve prediction accuracy and interpretability, especially in datasets with many predictors. For more detailed information, you can refer to the works by Friedman, Hastie, and Tibshirani (2010) and Hastie, Tibshirani, and Friedman (2009).

3. Random Forests (RF)

Random Forests (RF) is an ensemble learning method particularly suited for health economics due to its ability to handle complex datasets with interactions and non-linear relationships between variables, such as patient characteristics, provider attributes, and healthcare outcomes. RF operates by combining multiple decision trees to enhance predictive accuracy and mitigate overfitting, a common challenge in health economics datasets (Breiman, 2001).

The RF algorithm employs a bootstrap procedure to generate multiple subsets of the original dataset, with each subset used to train an individual decision tree. At each split in a tree, a random subset of predictors is evaluated to determine the optimal split, introducing additional randomness and fostering diversity among the trees. This process, known as random feature selection, is instrumental in capturing the intricate relationships within healthcare data. The final prediction is derived by aggregating the predictions from all trees, typically through majority voting for classification tasks or averaging for regression tasks. This aggregation step is pivotal as it reduces variance and bolsters the model's robustness, making it well-suited for the variability often encountered in health economics data (Hastie et al., 2009).

One of the standout advantages of RF in health economics is its capacity to automatically identify and incorporate interactions and non-linear relationships between variables, eliminating the need for manual feature engineering. This feature is invaluable for exploring complex healthcare datasets and developing predictive models that can inform policy decisions and resource allocation. Furthermore, RF provides measures of variable importance, which can aid in understanding the key drivers of healthcare outcomes and costs, thereby facilitating targeted interventions (Breiman, 2001).

Hyperparameter tuning is an essential component of RF, involving the adjustment of various parameters to optimize model performance. The number of trees in the forest is a critical hyperparameter; increasing the number of trees generally enhances the model's stability and accuracy. Another important hyperparameter is the number of predictors considered for splitting at each node, which influences the diversity of the trees. Splitting rules, such as Gini impurity for classification or mean squared error for regression, dictate the criteria for splitting nodes. The minimal node size, which specifies the minimum number of observations required in a terminal node, is also crucial for controlling the model's complexity and preventing overfitting (Hastie et al., 2009).

4. Gradient Boosting Machines (GBM)

Gradient Boosting Machines (GBM) is another ensemble learning method that builds trees sequentially, with each new tree aiming to correct the errors of the previous trees. This sequential learning approach makes GBM particularly effective for improving predictive performance in health economics, where accurate predictions are crucial for informed decision-making (Friedman, 2001). GBM employs gradient descent to minimize the loss function, iteratively refining the model by fitting each new tree to the residuals (errors) of the combined ensemble of all previous trees. This process allows the model to learn from its mistakes and gradually improve its predictions, making it well-suited for the complex and often noisy data encountered in health economics.

GBM's ability to handle complex interactions and non-linear relationships, similar to RF, is a significant advantage in health economics. However, GBM often achieves higher predictive accuracy than RF due to its sequential learning approach, making it a valuable tool for predicting healthcare outcomes, costs, and utilization patterns. GBM can be applied to both classification and regression tasks and can accommodate different types of loss functions, offering flexibility in modeling various healthcare phenomena (Chen & Guestrin, 2016).

Hyperparameter tuning is equally important in GBM, where several parameters need to be adjusted to optimize model performance. The number of trees is a key hyperparameter, determining the number of sequential trees built by the model. The interaction depth, which defines the maximum depth of each tree, controls the model's complexity and aids in capturing interactions between variables. Shrinkage, also known as the learning rate, is a factor that scales the contribution of each tree, helping to prevent overfitting by slowing down the learning process. The minimal node size, which specifies the minimum number of observations required in a terminal node, is another crucial hyperparameter for controlling the model's complexity (Friedman, 2001).

However, GBM is more computationally intensive and time-consuming, especially for hyperparameter tuning, compared to RF. This is due to the sequential learning process, which requires fitting each new tree to the residuals of the previous trees, a computationally expensive task. Additionally, GBM is prone to overfitting if not properly tuned, making it essential to carefully adjust the hyperparameters and validate the model's performance on a separate dataset. Despite these challenges, GBM's high predictive accuracy makes it a valuable tool for health economics research and applications (Chen & Guestrin, 2016).

Both RF and GBM are powerful tools for health economists, offering robust and flexible solutions for predictive modeling tasks. Their ability to handle complex datasets and capture intricate relationships makes them invaluable for informing policy decisions, resource allocation, and intervention strategies in healthcare.

References:

- Breiman, L. (2001). "Random Forests." *Machine Learning*, 45(1), 5-32.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
- Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, 29(5), 1189-1232.
- Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1–22.

- Chen, T., & Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Appendix 3 additional descriptives

Table A1 Total number of claims per year

Patient group	2016	2017	2018	2019	2020
Cataract	12297	11840	12448	11884	9451
ENT	18673	18743	19856	20987	13291
Intestinal cancer	3658	3463	3649	2980	2394

Table A2 Market shares per hospital in the pre- and post-periods for the three patient groups

	Cataract		**ENT**		**Intestinal cancer**	
Competitor number	Pre	Post	Pre	Post	Pre	Post
Slotervaart	5.8%	0.0%	7.4%	0.0%	10.8%	0.0%
Academic 1	0.0%	0.0%	3.9%	3.2%	1.9%	2.6%
Academic 2	0.0%	0.0%	4.3%	3.2%	2.8%	2.6%
Teaching 1	27.6%	22.5%	30.5%	32.1%	22.6%	27.0%
Top-clinical 1	3.7%	4.0%	0.0%	0.0%	7.4%	11.9%
General 1	5.2%	2.2%	7.1%	2.9%	5.1%	0.0%
General 2	0.0%	0.0%	0.0%	0.0%	1.6%	4.5%
General 3	6.3%	7.2%	9.6%	11.5%	5.6%	3.3%
General 4	0.0%	0.0%	10.3%	12.6%	7.9%	7.1%
General 5	0.0%	0.0%	10.0%	11.2%	9.5%	11.3%
ITC 1	15.2%	24.5%	0.0%	0.0%	8.0%	14.4%
ITC 2	0.0%	0.0%	0.0%	0.0%	5.3%	0.0%
ITC 3	0.0%	0.0%	0.0%	0.0%	2.1%	0.0%
ITC 4	11.7%	12.5%	0.0%	0.0%	0.0%	0.0%
ITC 5	0.0%	0.0%	0.0%	0.0%	1.3%	1.2%
ITC 6	0.0%	0.0%	10.7%	12.0%	0.0%	0.0%
ITC 7	0.0%	0.0%	3.1%	5.9%	0.0%	0.0%
ITC 8	0.0%	0.0%	2.9%	5.5%	0.0%	0.0%
ITC 9	7.0%	10.5%	0.0%	0.0%	0.0%	0.0%
ITC 10	7.0%	7.9%	0.0%	0.0%	0.0%	0.0%
ITC 11	0.0%	0.0%	0.0%	0.0%	5.9%	10.1%
ITC 12	10.4%	8.7%	0.0%	0.0%	0.0%	0.0%
Other 1	0.0%	0.0%	0.0%	0.0%	2.4%	4.2%

Appendix 4 shrunk LASSO coefficients

Table A3 Shrunk LASSO coefficients

	Cataract	**ENT**	**Intestinal cancer**
Variable			
Distance (km)			
Linear Term	-0.162	-0.162	-0.082
Quadratic Term	0.001	0.001	0
Academic*Distance		0.075	0.122
Top-clinical*Distance	0		0.008
Teaching*Distance	0.019	0.039	0.066
ITC*Distance	0.077	0.005	0.1
Other*Distance			0.097
Age*Distance	0	0	-0.001
Female*Distance	0	0	0
Time			
Linear Term	-0.039	-0.052	-0.048
Quadratic Term	0	0	0
Academic*Time		0	-0.026
Top-clinical*Time	0.091		0.016
Teaching*Time	0.082	0.068	0.038
ITC*Time	0	0.01	0.016
Other*Time			0.055
Age*Time	0	0	0
Female*Time	-0.001	0	0

Appendix 5 Sensitivity analyses

Using multiple years of data

Figure A2 Mean Absolute difference between Estimated and actual diversion ratios for three patient groups using January 2014—July 1 2018 as the training data and January 2019—December 2022 as the test set

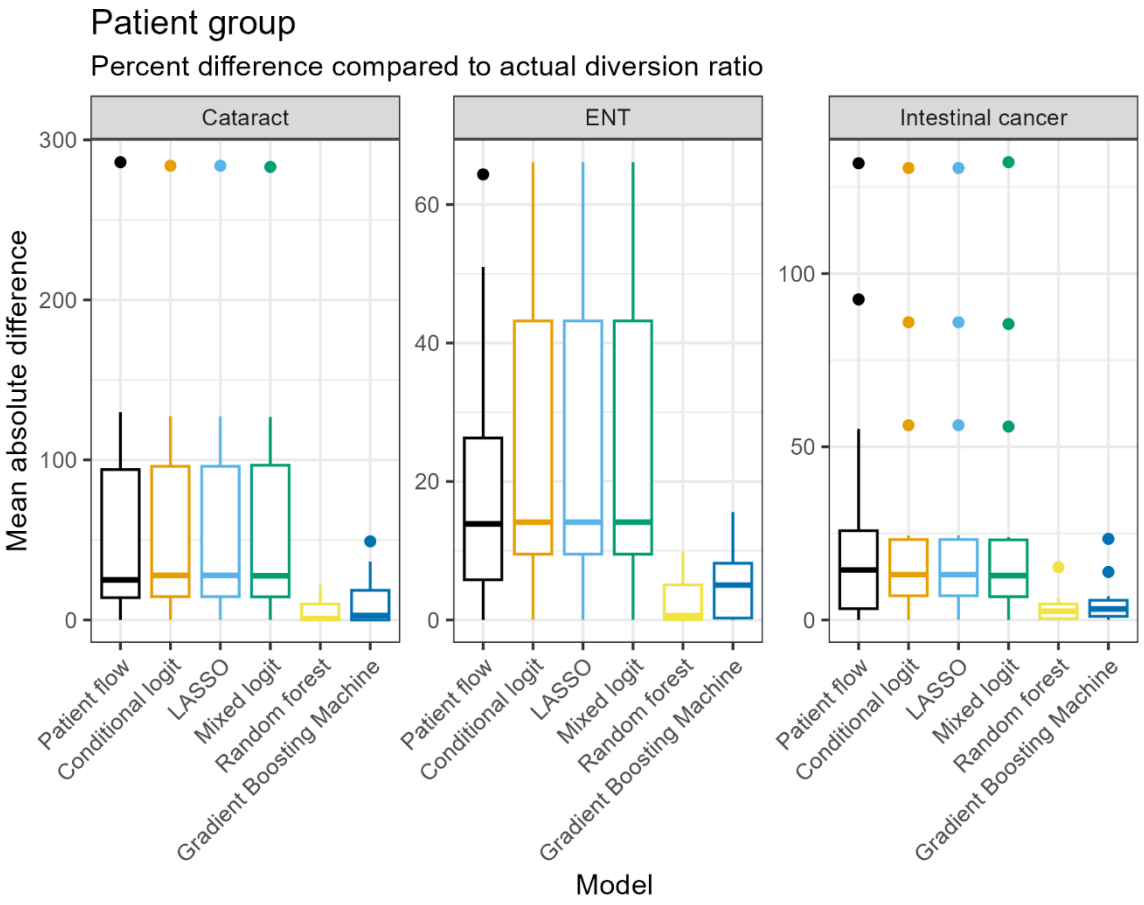
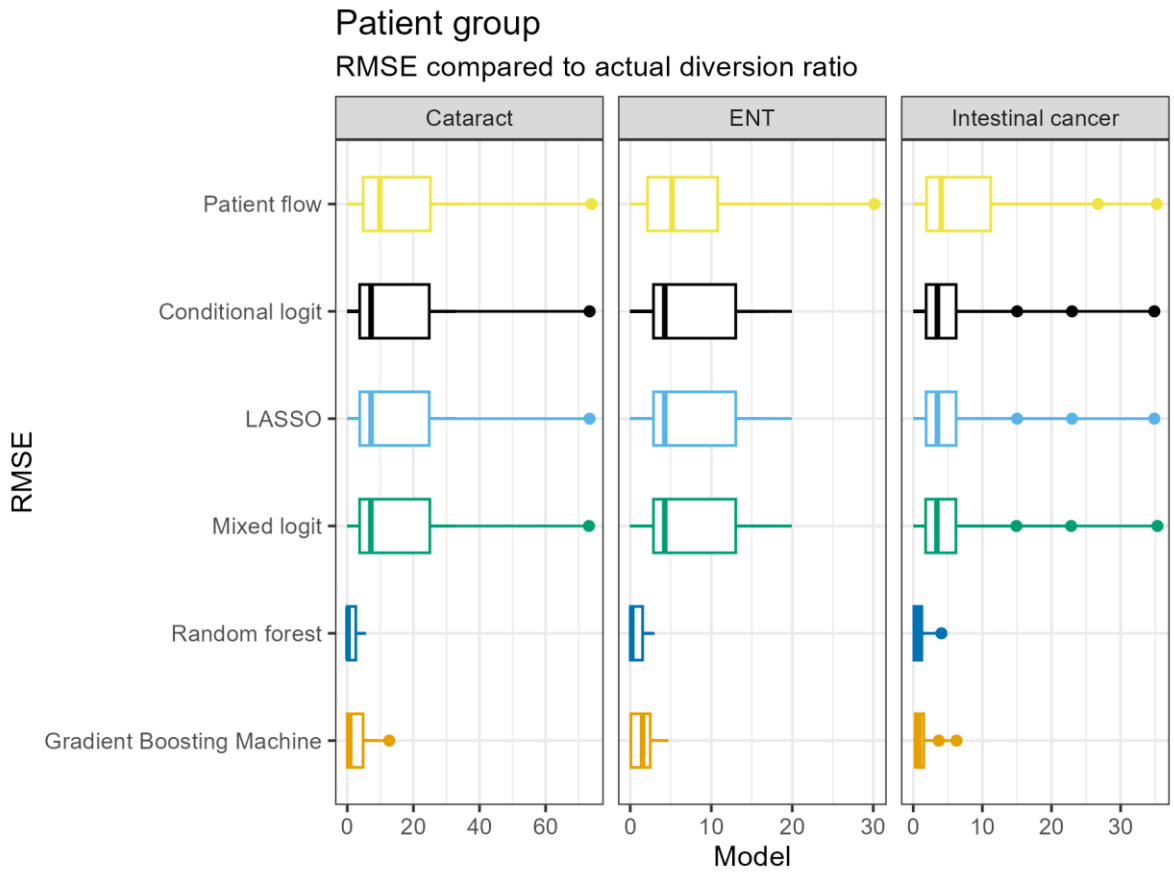


Figure A3. Percent improvement in RMSE versus the observed diversion ratios for three patient groups using January 2014—July 1 2018 as the training data and January 2019—July 1 2022 as the test set



Nearby vs. far away patients

Figure A4 Mean Absolute difference between Estimated and actual diversion ratios for three patient groups after splitting the group into “nearby” and “faraway”

Nearby *Far away*

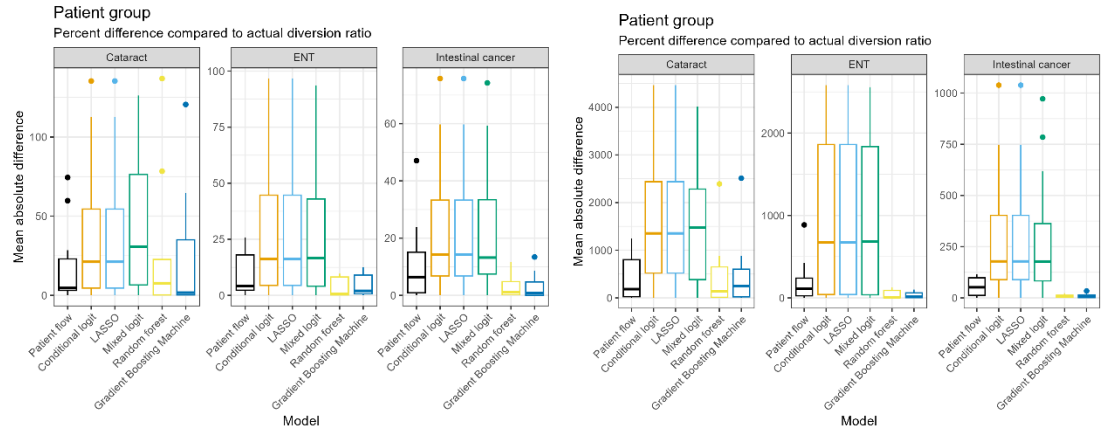
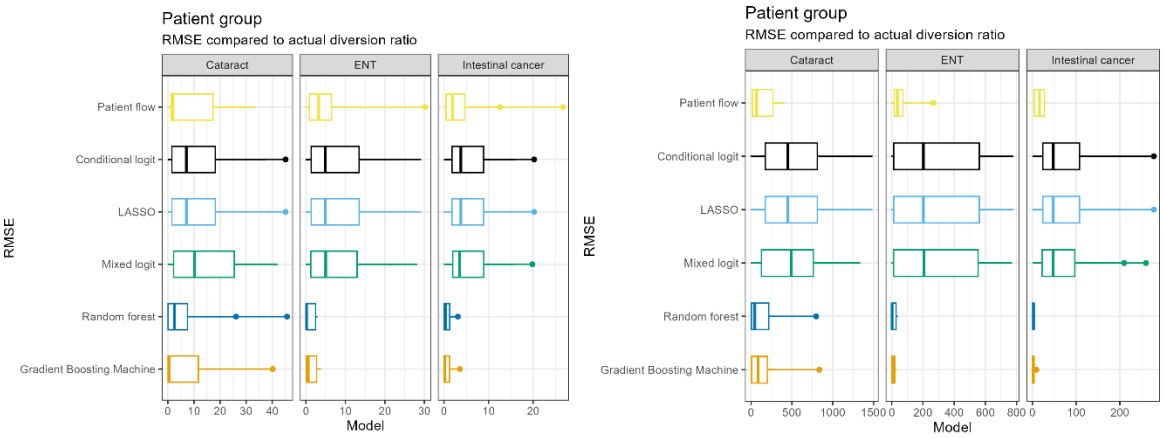


Figure A5. Percent improvement in RMSE versus the observed diversion ratios for three patient groups after splitting the group into “nearby” and “faraway”

Nearby Far away



All patient groups

Figure A6 MAD plot for all patient groups

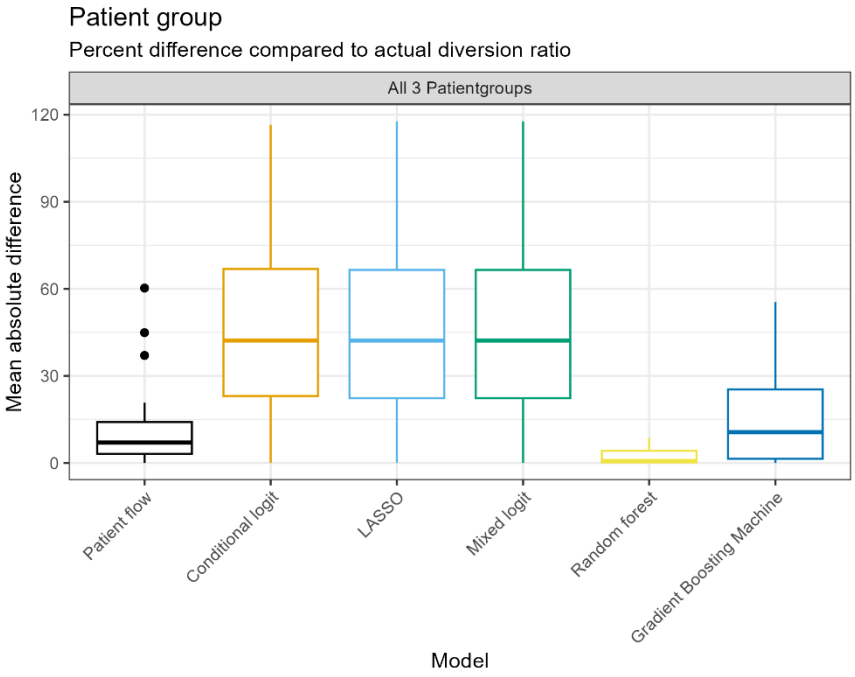
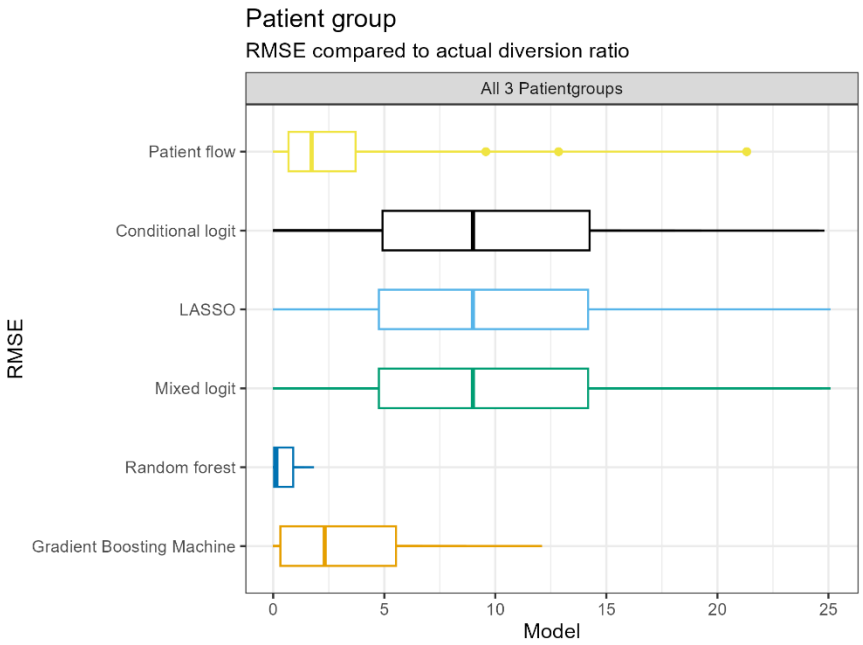


Figure A7 RMSE



Appendix 6 Results per hospital

Figure A8 Cleveland plot per hospital MAD

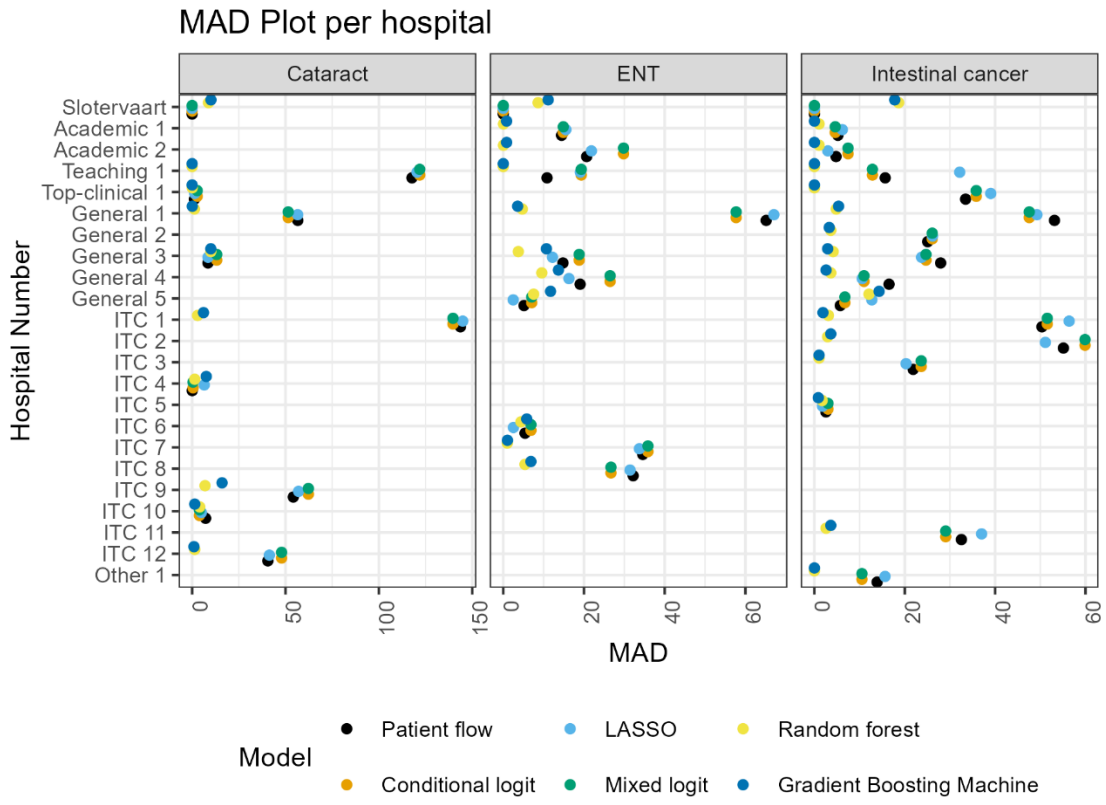
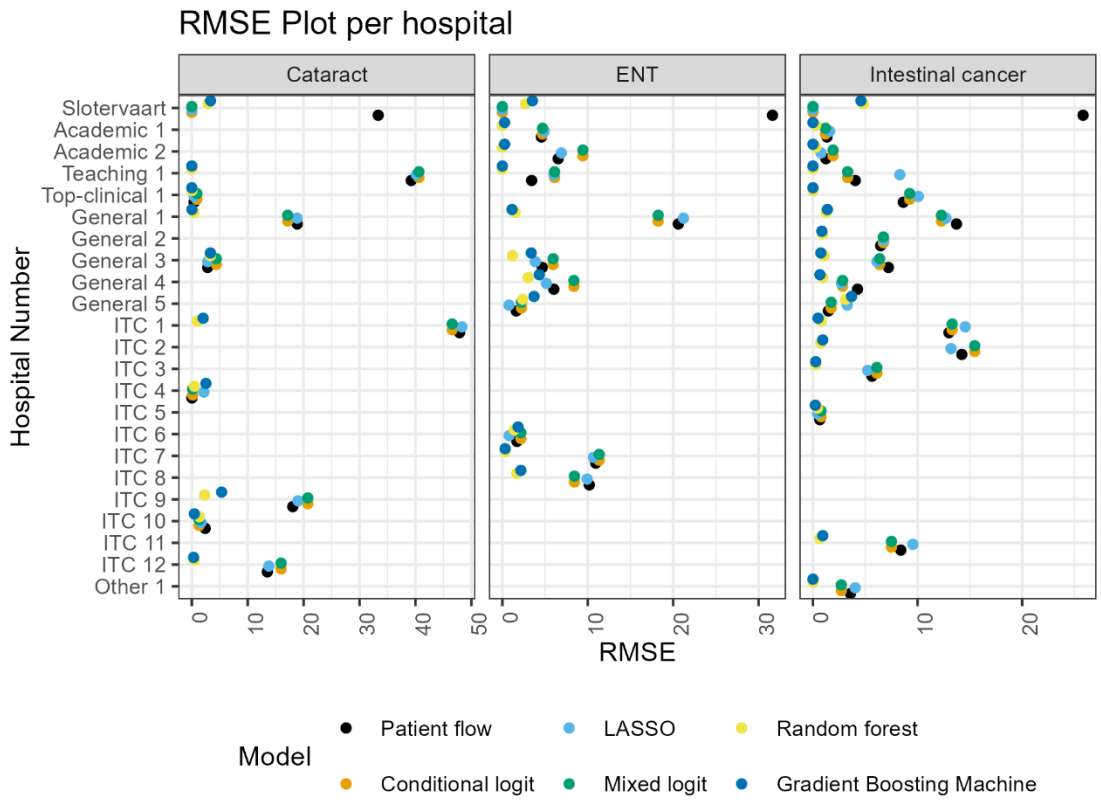


Figure A9 RMSE plot per hospital


RMSE Plot per hospital
Cataract
ENT
Intestinal cancer
Hospital Number
Slotervaart
Academic 1
Academic 2
Teaching 1
Top-clinical 1
General 1
General 2
General 3
General 4
General 5
ITC 1
ITC 2
ITC 3
ITC 4
ITC 5
ITC 6
ITC 7
ITC 8
ITC 9
ITC 10
ITC 11
ITC 12
Other 1
0
10
20
30
40
50
RMSE
Model
Patient flow
LASSO
Random forest
Conditional logit
Mixed logit
Gradient Boosting Machine