Title

Machine learning models to predict length of stay and discharge destination in complex head and neck surgery

Permalink
https://escholarship.org/uc/item/39g5t9w5

Journal
Head & Neck, 43(3)

ISSN
1043-3074

Authors
Goshtasbi, Khodayar
Yasaka, Tyler M
Zandi-Toghani, Mehdi
et al.

Publication Date
2021-03-01

DOI
10.1002/hed.26528

Peer reviewed

Published in final edited form as:
Head Neck. 2021 March ; 43(3): 788–797. doi:10.1002/hed.26528.

Machine Learning Models to Predict Length of Stay and Discharge Destination in Complex Head and Neck Surgery

Khodayar Goshtasbi, MS[1,*], **Tyler M. Yasaka, BS**[1,*], **Mehdi Zandi-Toghani, MS**[1], **Hamid R. Djalilian, MD**[1,2], **William B. Armstrong, MD**[1], **Tjoson Tjoa, MD**[1], **Yarah M. Haidar, MD**[1], **Mehdi Abouzari, MD, PhD**[1]

[1]Department of Otolaryngology–Head and Neck Surgery, University of California, Irvine, USA

[2]Department of Biomedical Engineering, University of California, Irvine, USA

Abstract

Background: This study develops machine learning (ML) algorithms that use preoperative-only features to predict discharge-to-nonhome-facility (DNHF) and length-of-stay (LOS) following complex head and neck surgeries.

Methods: Patients undergoing laryngectomy or composite tissue excision followed by free tissue transfer were extracted from the 2005–2017 NSQIP database.

Results: Among the 2786 included patients, DNHF and mean LOS were 421 (15.1%) and 11.7±8.8 days. Four classification models for predicting DNHF with high specificities (range, 0.80–0.84) were developed. The generalized linear and gradient boosting machine models performed best with receiver operating characteristic (ROC), accuracy, and negative predictive value (NPV) of 0.72–0.73, 0.75–0.76, and 0.88–0.89. Four regression models for predicting LOS in days were developed, where all performed similarly with mean-absolute-error and root-mean-squared-errors of 3.95–3.98 and 5.14–5.16. Both models were developed into an encrypted web-based interface: https://uci-ent.shinyapps.io/head-neck/.

Conclusion: Novel and proof-of-concept ML models to predict DNHF and LOS were developed and published as web-based interfaces.

Keywords

Machine learning; artificial intelligence; prediction; length of stay; discharge

Introduction

Machine learning (ML) is an analytical application of artificial intelligence with the ability to "learn" from new information, without explicit directions or programming, for self-improvement. In contrast to the majority of predictive models in clinical literature which are

Corresponding Author: Mehdi Abouzari, MD, PhD, Department of Otolaryngology–Head and Neck Surgery, University of California Irvine, 333 City Blvd. West, Suite 525, Orange, CA 92868, Phone: (714) 509-6096, Fax: (714) 456-5747, mabouzar@hs.uci.edu.

[*]These authors contributed equally to this manuscript.

Conflicts of Interest: None

associative and reliant on direct coefficient-outcome relationships, ML predictive models may potentially improve outcomes by executing complex, hidden, and non-linear computations.[1] There has been an emerging interest in utilizing ML in the medical field for genomic[2] or imaging classifications,[3] as well as predicting disease prognosis[4] and treatment complications.[5] Most recently in the otolaryngology literature, authors have reported ML algorithms for detecting pharyngeal cancer,[6] estimating head and neck (HN) squamous cell carcinoma prognosis,[7] and predicting complications following HN microvascular free tissue transfer.[8] To date, no study has used ML to predict post-operative length of stay (LOS) or discharge to nonhome facility (DNHF) following complex HN surgeries. Furthermore, although the majority of the clinical ML studies focus on common algorithmic models such as artificial neural network (ANN) and generalized linear model (GLM), there exist other complex ML models that may provide more superior predictive capabilities.[9, 10]

Our current understanding of important clinical prognosticators in HN surgeries have benefited immensely from traditional statistical models. However, with the complexity and interactivity of the vast number of pre-operative clinical variables, and the advent of large publicly available databases, investigators can now attempt developing sophisticated models that factor in many input data to predict an outcome. Type and complexity of ML models are dependent on the application and training dataset, and though it may sound counterintuitive, more complexity is not always better.[11] An appropriately constructed and tested ML model can be a valuable tool in the HN field by identifying at-risk patients and helping providers appropriately pre-plan operations. In this manuscript, we will construct GLM, ANN, support vector machine (SVM), random forest (RF), and gradient boosting machine (GBM) models to predict LOS and DNHF, which are important outcome variables in the field of HN surgery. Notably, longer LOS can be associated with increased morbidity (e.g., infections)[12] and higher healthcare cost,[13] and nonhome facility placement can associate with insurance difficulties and delayed discharge.[14, 15] As such, this proof-of-concept study aims to construct different ML models that predict LOS and DNHF following complex HN surgeries, compare these algorithms' performance to each other, and publish the best-performing models as public web-based interfaces for the readership.

Materials and Methods

Patient Population

The 2005–2017 American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) database, which collects 30-day morbidity and mortality information for various operations, was retrospectively reviewed for data collection. Given the de-identified and publicly available nature of this database, this study was exempted from Institutional Review Board approval. Complex head and neck surgeries were defined as laryngectomy or composite tissue excision followed by free tissue transfer, modeled after a recent NSQIP study by Lebo and colleagues.[16] The following current procedural terminology (CPT) codes were used to collect patients undergoing laryngectomy: 31360, 31365, 31368, 31390, and 31395. For the other group of patients, inclusion criteria required two linked surgeries: 1) head and neck mucosal or composite resection with CPTs including 21034, 21044, 21045, 21047, 31230, 31225, 40814, 40816, 41116, 41120, 41130, 41135, 41140, 41145, 42120, 41150, 41153, 41155,42845, 42894, and 2) free tissue transfer with CPTs including 15756 15757, 15758, 20955, 20956, 20962, 20969, and 20970.

Covariates and Accounting for Missing Values

The detailed description of the NSQIP variables are found in the user guide for the 2017 ACS-NSQIP Participant Use Data File. The American Society of Anesthesiologists (ASA) score which measures pre-operative comorbidities and overall health was binarized as low (class 1–2) and high (class 3–4) ASA class. DNHF included skilled care or unskilled facility, rehabilitation center, or separate acute care center. Furthermore, LOS was defined as days from operation to discharge. The NSQIP data, similar to other national databases, suffers from a considerable amount of missing values. As such, input variables with more than 25% missing values (e.g., albumin and prothrombin time levels) were excluded from this study. Due to the importance of properly handling missing data, we used the *missForest* package in the R statistical programming language,[17] which imputed missing values in a manner less prone to bias compared to alternative methods of handling missing values.[18] To determine which factors would be included for DNHF or LOS models, univariate analysis (e.g., chi-square, independent *t*-test, and Pearson correlation) was performed to evaluate the association between the pre-operative input and outcome variable, and those with *P* value <0.2 or deemed clinically important were included.

Predictive Modeling and Statistical Analysis

The dataset was randomly stratified into a training and testing set using an 80:20 ratio separately for the DNHF and LOS models. The training sets were used to train the algorithms, and free parameters were adjusted according to results from the training set's cross-validation. Hyperparametric optimization was achieved using random search methods.[19] Our classification models were configured to output probabilities for each prediction, with predictions above a given probability (the "threshold") classified as positive outcomes.[20] Class imbalance (when one outcome label is significantly less prevalent than another outcome label) was addressed using weighed distance function, which is arguably more effective than commonly used over- or under-sampling methods.[21] After optimizing the classification models for the receiver operating characteristic area under the curve (ROC-AUC), the classification thresholds were adjusted to target specificities of approximately 80%. Models were evaluated via internal validation, where each model was trained on the training partition and predictions on the test set were evaluated using multiple performance metrics. The reported performance metrics are based on averages of twenty trials to control for bias introduced by randomness in the models. All statistical analyses, including ML development/testing and figure generation, were performed using R version 3.6.3 (The R Foundation for Statistical Computing, Vienna, Austria) via RStudio version 1.1.463 (RStudio, Boston, MA).

The two models had slightly different cohorts depending on exclusion criteria: For DNHF classification models, patients with unknown discharge information were excluded (N=386), and for LOS regression models, patients with unknown or >30-day discharge were excluded (N=119). A total of four classification models were trained and tested to predict DNHF: GLM, ANN, RF, and GBM. Since DNHF was a binary outcome variable, the performance of these models was assessed using sensitivity, specificity, ROC-AUC, positive predictive value (PPV), negative predictive value (NPV), and accuracy. Four regression models were trained and tested for predicting days of post-operative LOS: GLM, ANN, RF, and GBM. Since LOS was treated as a continuous outcome variable, the performance of the regression models was assessed using mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and root mean squared logarithmic error (RMSLE). To evaluate how the regression models performed according to the extent of LOS, patients were divided into quartiles and the relative prediction performances (assessed by RMSE) were compared accordingly.

Results

Patient Characteristics

A total of 2,786 patients were extracted for further analysis, 750 (26.9%) of which were female and 1870 (67.1%) were white. Mean age, body mass index (BMI), and operation time were 62.1 ± 11.7 years, 25.3 ± 6.3, and 485.5 ± 196.6 minutes, respectively.

Discharge to Nonhome Facility

A total of 2,400 patients were included for these classification models, which included 421 (15.1%) DNHF patients. Variables included for DNHF predictive models were age, BMI, gender, race, procedure type, diabetes, smoking, dyspnea, functional status, ventilator dependency, history of chronic obstructive pulmonary disease or congestive heart failure, hypertension, ASA class, chronic steroid use, emergency or elective surgery, pre-operative sepsis or transfusion, history of wound infection, and pre-operative sodium, BUN, creatinine, WBC, HCT, and platelet levels (Table 1). Four models were constructed to predict DNHF, and their performance on the testing set are demonstrated in Table 2. Overall, with the optimized specificity of approximately 0.80–0.84 for all models, accuracy and sensitivity ranged between 0.73–0.76 and 0.39–0.53, respectively. GLM and GBM had the highest ROC's of 0.72–0.73 (Figure 1). Since the algorithms' designated classification thresholds are directly related to the resulting sensitivity and specificity, the associations between these two metrics and the thresholds are depicted in Figure 2. This demonstrates that although the majority of our reported metrics were dependent on the optimization of ~0.80 specificity, adjusting the threshold to a value corresponding to lower specificity may raise the sensitivity; for instance, in the GBM model we developed, a balance of both sensitivity and specificity at around 0.65 could be achieved simply by using a lower classification threshold (Figure 2B).

Length of Stay

A total of 2,667 patients were included for the regression-based (continuous) LOS predictive models, with a mean LOS following complex HN surgeries of 10.4 ± 5.5 days. The input variables included age (P<0.01), BMI (P=0.02), gender (P=0.52), race (P<0.01), procedure type (P=0.08), elective surgery (P<0.01), diabetes (P=0.05), smoking (P=0.02), dyspnea (P=0.01), hypertension (P=0.08), dependency functional status(P<0.01), history of congestive heart failure (P=0.05), disseminated cancer (P<0.01), history of wound infection (P=0.02), pre-operative sepsis (P=0.09) or transfusion (P=0.10), ASA class (P=0.06), and pre-operative sodium (*P*<0.01), WBC (*P*<0.01), and HCT (*P*<0.01), with the *P* values representing their association with LOS on independent *t*-test (for categorical variables such as gender) or correlation analysis (for continuous variables such as age). Four models were constructed to predict LOS, and their performance on the testing set are demonstrated in Table 3. The performances of these models were compared according to different extents of LOS, demonstrating that these models predicted LOS better when the actual LOS was >8 days (Figure 3). For both DNHF and LOS GLM models, the average contribution of each feature in the prediction of the outcomes were calculated over 20 iterations, and those with the largest contributions are demonstrated in Table 4.

Predictive Modeling Interface

The best-performing algorithms for predicting DNHF and LOS were developed into an encrypted web-based interface which can be accessed at https://uci-ent.shinyapps.io/head-neck/.

Discussion

Discharge destination and prolonged hospitalization are important post-operative outcomes in complex surgeries including head and neck operations. Although ML models to predict these in neurological or orthopedic surgeries have been developed using the NSQIP database,[9, 22–25] there is a paucity of similar investigations in the otolaryngology literature. To our knowledge, this is the first manuscript that develops proof-of-concept ML algorithms to pre-operatively predict DNHF and LOS following complex HN surgeries, and among the minority of studies to publish the developed ML models as a public interface for simulation and further examination by the readership. Such predictive models can identify patients at risk of DNHF or prolonged LOS, so their pre-, peri-, and post-operative planning can be evaluated with caution.

Although Panwar et al.[26] and White et al.,[27] have previously used traditional statistics to investigate LOS and DNHF outcomes in NSQIP's HN patients, this study's novelty lies in its ML approach (both classification and regression models) and publishing these as publicly available web-based interfaces where new input data can produce real-time predictions. Despite their limitations, these ML models can serve as proof-of-concept applications that can be improved with future multi-institutional investigations. Additionally, this manuscript further suggests factors that may heavily contribute to the studied outcomes as summarized in Table 4, including pre-operative transfusion, dependent functional status, and history of congestive heart failure, wound infection, smoking, or diabetes. The published web-based model interfaces are not FDA-approved or ready for real-life clinical decision-making, but they serve as important proof-of-concept demonstrations of how this field of research can hypothetically provide practical clinical value in the future.

Utilizing the NSQIP dataset and only pre-operative features, our DNHF classification models demonstrated that GLM and GBM models performed well with ROC-AUC of 0.72 and 0.73, followed by ANN and RF models with AUC-ROC of 0.67 and 0.66, respectively. These models were classified to optimize specificity to approximately 0.80, resulting in high NPVs ranging 0.86–0.89. Moreover, we successfully built four regression ML models that performed similarly in predicting LOS (in days) with MAE and RMSE of 3.9–4.0 and 5.1–5.2, respectively. The models were trained using k-fold cross-validation.[28] Although the present results lack external validation, we have published both models via an encrypted open-access interface for the readership as a means of potential future external validation. The different performance of our predictive models compared to other ML publications[9, 22] may stem from several reasons[11]: 1) feature selection and the algorithm's attributed weight of each input variable; 2) finding optimal free parameter values used for feature transformation and class-based prediction; 3) balancing the trade-off between model complexity and generalizability to novel cases without overfitting;[29] and finally 4) the inherit associative strengths of the features in predicting the outcome of interest. To achieve optimal predictive performance, it is imperative that multiple models and parametric configurations be developed and that a robust training method (such as k-fold cross-validation) be utilized to minimize the risk of overfitting. Further, it is worth emphasizing that the performance and validity of a ML model depends on the nature of the dataset and the degree of underlying direct or hidden relationships between the features and outcome.[30–32]

All of the developed models were supervised ML constructs. GLM algorithms which use traditional regression mathematical models have been most frequently utilized for constructing predictive models in the medical literature.[33] Based on statistical learning theory introduced by Vapnik,[34] ANNs are gross neuron-by-neuron simulations of the human brain with finite numbers of layers, nodes, interconnections, and weighted variables.[35] RF is a type of decision tree algorithm where different data sample bootstraps are utilized for creating each tree, where the number of trees and model predictors positively correlate.[36] Another form of decision tree model is GBM, which fits weak learner decision tress to a regression model,[37] with the advantage of being highly adaptable and interpretable.[38] The fact that GLM and GBM models were the most ideal in predicting DNHF in this cohort suggests that model performance is not always correlated with its complexity, but also heavily dependent on the dataset and nature of input-output relationships. Although previous papers in orthopedic and cardiology literature have developed classification models for LOS prediction (e.g., short *vs.* long LOS),[39, 40] our proof-of-concept regression model with numeric outputs for LOS (in days) was more novel to the literature, paving path for future studies to develop such algorithms.

As we demonstrated in Figure 2, adjusting the classification threshold values determines the balance between sensitivity and specificity of the predicted binary outcomes, which is an important theme when considering the model's practicality and intended application. Of note, changing the threshold does not change the underlying model; rather, the model outputs a probability for each prediction, after which the threshold is applied such that probabilities above the threshold will be classified as positive (e.g., DNHF). Our DNHF results prioritizing specificity over sensitivity expressed a preference for having high confidence in positive predictions, as opposed to preferring to detect a greater proportion of positive outcomes (which would be prone to more false alarms). It is also important to emphasize the time-point at which the model is intended to be utilized. Our designed *pre-operative* models precluded the use of intra- and post-operative variables which could have strong associations with the outcome. This distinction is further demonstrated in a study predicting readmission following spine surgery, where the aggregate ML models' ROC decreased from 0.81 to 0.58 when only including pre-operative characteristics.[41] When constructing a ML model, it is crucial to consider the applicability of such a model. If, for example, a model intended to predict surgical outcomes for pre-operative planning includes intra- or post-operative variables, the model may fail to provide any practical benefit.

Although the developed models utilized variables which underwent one round of inclusion/exclusion analysis (per Table 1) before being incorporated in the algorithms, a few variables could theoretically be removed without significantly changing the performances. One way to address this is to investigate which variables had the least amount of weight for a given model's outcome predictions (e.g., Table 4) and remove variables in a stepwise manner. Although we will continue improving the published models and their online interfaces by continuously adding subjects and re-examining which variables continue to have insignificant effects, the current interface at its proof-of-concept stage could benefit from including all the possibly contributing variables. One reason for avoiding over-simplification of variable numbers at this proof-of-concept stage is so the readership can also investigate the influence of each variable in real time. For instance, data of a hypothetical patient could be entered in the algorithm, then certain variables (e.g., pre-operative BUN or dyspnea) could be changed to see if/how the outcome probability changes. This is especially an advantage of machine learning over traditional statistical models due to the former's ability to deduct potentially complex, non-linear, and hidden relationship with each variable in regard to others (e.g., if dyspnea has a higher association with a certain outcome depending on the patient's age, BMI, and ASA combination).

Discussing the limitations of this study will be important for future progress of this field. First, this study utilized retrospective data from a de-identified national database, which is prone to missing values, miscoding, or inherent biases. Although using the *missForest* R package is a rigorous technique for missing value imputation,[18] this approach of running a ML model on a partially ML-generated dataset can introduce some systemic bias. Second, the database did not include some important clinical or socioeconomic variables which may associate with DNHF or LOS. Although this limits the scope of the models, we believe that the presented models can serve as proof-of-concept applications warranting future large-scale studies using more comprehensive datasets. Also, we chose not to include intra- and post-operative variables (e.g., operation time, transfusion, post-operative complications, etc.) even though they may strongly associate with DNHF or LOS. For instance, LOS itself can associate with DNHF.[42] This was because our models were intended to guide clinicians at a pre-operative standpoint to provide a window of opportunity for appropriate planning or arrangements for risk adjustment. While such models cannot serve as decision-making bodies, they may serve as supplementary tools in screening for patients with relatively higher probabilities of DNHF or prolonged LOS, thus providing opportunities for early planning and risk management. Although an aforementioned study[41] did not surpass a 60% ROC-AUC using pre-operative-only characteristics compared to an 81% ROC-AUC when also incorporating post-operative characteristics, our proof-of-concept study demonstrated that a pre-operative-only study (which is more practical and valuable as argued) can reach an acceptable (>70% ROC-AUC) success rate. This warrants future investigations to similarly build machine learning models while considering model timeline and practicality (e.g., providing opportunities for clinical actions or precautions) and cautiously start incorporating variables from different timelines. Future studies should also investigate other national databases (e.g., National Cancer Database, Surveillance Epidemiology and End Results Database) or multi-institutional cohorts which can potentially provide many more pre-operative or long-term post-operative variables useful for building these models. As such, following this proof-of-concept study, our next phase of investigations will utilize data with more available variables including those shown to influence treatment decision-making and outcomes such as various socioeconomics[43–45] and staging/grading information.[46–48] Institutional studies could also factor in important features not captured by national databases but nevertheless influential towards treatment outcomes such as family support, physician discussions, physician peri/post-operative protocols and wound care, tracheostomies/gastrostomies, employment, and type and location of the treatment facility, among others. Third, although this study contains internal validation by partitioning of the cohort into training and testing sets, there was no external validation which would require testing outside patients. To address this, we have published our algorithms via an encrypted online interface and encourage clinicians to test hypothetical situations. Fourth, patients undergoing laryngectomy or composite tissue excision followed by free tissue transfer were combined similar to a previous study regarding complex head and neck surgery patients.[16] Following this proof-of-concept investigation, future studies using multi-institutional cohorts or other national databases will allow utilization of more homogenous surgical cohorts with adequate reliability and number of subjects and variables. Finally, while our models were carefully tuned to maximize predictive performance, they were limited by the extent of inherent relationships between the features and outcomes. Despite these limitations, developing such assisting tools that may predict DNHF or prolonged LOS with high specificity has the potential to allow clinicians to identify at-risk patients and adjust pre- and post-operative planning accordingly. Future research may evaluate our published predictive models on external data sets or improve upon them by utilizing more comprehensive data sets or integrating institutional data with otolaryngology-specific features.

Conclusion

In this proof-of-concept manuscript, ML algorithms were developed to predict DNHF and LOS following complex head and neck surgeries, and the best-performing models were published as an encrypted public interface for the readership (https://uci-ent.shinyapps.io/head-neck/). Our best performing classification model in predicting DNHF was GLM, while all four regression models, namely GLM, ANN, RF, and GBM, performed comparably for predicting LOS. The discussion of features and threshold determinations, rigor of ML training and validation, and tradeoffs between sensitivity-specificity or complexity-generalizability are important ML topics that warrant continuous investigations.

Financial Disclosure:

Mehdi Abouzari is supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant TL1TR001415.

References

1. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216. [PubMed: 27682033]
2. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16(6):321–32. [PubMed: 25948244]
3. Nguyen LD, Lin D, Lin Z, Cao J. Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation. 2018 IEEE International Symposium on Circuits and Systems (ISCAS); pp. 1–5. IEEE, 2018.
4. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13:8–17. [PubMed: 25750696]
5. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, Motaei A, Madkour M, Pardalos PM, Lipori G, Hogan WR, Efron P, Moore F, Moldawer L, DWang DZ, Hobson CE, Rashidi P, Li X, Momcilovic M. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. Ann Surg. 2019;269(4):652–62. [PubMed: 29489489]
6. Tamashiro A, Yoshio T, Ishiyama A, Tsuchida T, Hijikata K, Yoshimizu S, Horiuchi Y, Hirasawa T, Seto A, Sasaki T, Fujisaki J, Tada T. Artificial‐intelligence‐based detection of pharyngeal cancer using convolutional neural networks. Dig Endosc. 2020 [Epub ahead of print]
7. Xing L, Zhang X, Guo M, Zhang X, Liu F. Application of Machine Learning in Developing a Novelty Five-Pseudogene Signature to Predict Prognosis of Head and Neck Squamous Cell Carcinoma: A New Aspect of “Junk Genes” in Biomedical Practice. DNA Cell Biol. 2020 [Epub ahead of print]
8. Formeister EJ, Baum R, Knott PD, Seth R, Ha P, Ryan W, El-Sayed I, George J, Larson A, Plonowska K, Heaton C. Machine Learning for Predicting Complications in Head and Neck Microvascular Free Tissue Transfer. Laryngoscope. 2020 [Epub ahead of print]
9. Goyal A, Ngufor C, Kerezoudis P, McCutcheon B, Storlie C, Bydon M. Can machine learning algorithms accurately predict discharge to nonhome facility and early unplanned readmissions following spinal fusion? Analysis of a national surgical registry. J Neurosurg Spine. 2019;31(4):568–78.
10. Karhade AV, Thio QC, Ogink PT, Shah AA, Bono CM, Oh KS, Saylor PJ, Schoenfeld AJ, Shin JH, Harris MB, Schwab J. Development of machine learning algorithms for prediction of 30-day mortality after surgery for spinal metastasis. Neurosurgery. 2019;85(1):E83–E91. [PubMed: 30476188]
11. Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920–30. [PubMed: 26572668]
12. Rosman M, Rachminov O, Segal O, Segal G. Prolonged patients’ In-Hospital Waiting Period after discharge eligibility is associated with increased risk of infection, morbidity and mortality: a retrospective cohort analysis. BMC Health Serv Res. 2015;15(1):246. [PubMed: 26108373]
13. Missios S, Bekelis K. Drivers of hospitalization cost after craniotomy for tumor resection: creation and validation of a predictive model. BMC Health Serv Res. 2015;15(1):85. [PubMed: 25756732]
14. Carey MR, Sheth H, Braithwaite RS. A prospective study of reasons for prolonged hospitalizations on a general medicine teaching service. J Gen Intern Med. 2005;20(2):108–15. [PubMed: 15836542]
15. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007;115(7):928–35. [PubMed: 17309939]
16. Lebo NL, Quimby AE, Caulley L, Thavorn K, Kekre N, Brode S, Johnson-Obaseki S. Surgical Site Infection Affects Length of Stay After Complex Head and Neck Procedures. Laryngoscope. 2020 [Epub ahead of print]
17. Stekhoven DJ. Using the missForest package. R package. 2011:1–11.
18. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112–8. [PubMed: 22039212]
19. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13(Feb):281–305.

20. Zou Q, Xie S, Lin Z, Wu M, Ju Y. Finding the best classification threshold in imbalanced classification. Big Data Research. 2016;5:2–8.

21. Kotsiantis S, Kanellopoulos D, Pintelas P. Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering. 2006;30(1):25–36.

22. Ogink PT, Karhade AV, Thio QC, Gormley WB, Oner FC, Verlaan JJ, Schwab JH.. Predicting discharge placement after elective surgery for lumbar spinal stenosis using machine learning methods. Eur Spine J. 2019;28(6):1433–40. [PubMed: 30941521]

23. Ogink PT, Karhade AV, Thio QC, Hershman SH, Cha TD, Bono CM, et al. Development of a machine learning algorithm predicting discharge placement after surgery for spondylolisthesis. Eur Spine J. 2019;28:1775–82. [PubMed: 30919114]

24. Biron DR, Sinha I, Kleiner JE, Aluthge DP, Goodman AD, Sarkar IN, et al. A Novel Machine Learning Model Developed to Assist in Patient Selection for Outpatient Total Shoulder Arthroplasty. J Am Acad Orthop Surg. 2019;28(13):e580–e585.

25. Muhlestein WE, Akagi DS, Davies JM, Chambless LB. Predicting inpatient length of stay after brain tumor surgery: Developing machine learning ensembles to improve predictive performance. Neurosurgery. 2019;85(3):384–93. [PubMed: 30113665]

26. Panwar A, Wang F, Lindau R, Militsakh O, Coughlin A, Smith R, et al. Prediction of discharge destination following laryngectomy. Otolaryngol Head Neck Surg. 2018;159(6):1006–11. [PubMed: 30126321]

27. White LJ, Zhang H, Strickland KF, El-Deiry MW, Patel MR, Wadsworth JT, et al. Factors associated with hospital length of stay following fibular free-tissue reconstruction of head and neck defects: assessment using the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) criteria. JAMA Otolaryngol Head Neck Surg. 2015;141(12):1052–8. [PubMed: 25905986]

28. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Statistics surveys. 2010;4:40–79.

29. Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in perforance evaluation. J Mach Learn Res. 2010;11(Jul):2079–107.

30. Riley P Three pitfalls to avoid in machine learning. Nature. 2019;572(7767):27–29. [PubMed: 31363197]

31. Brynjolfsson E, Mitchell T. What can machine learning do? Workforce implications. Science. 2017;358(6370):1530–4. [PubMed: 29269459]

32. Rowe M An introduction to machine learning for clinicians. Acad Med. 2019;94(10):1433–6. [PubMed: 31094727]

33. Hasan O, Meltzer DO, Shaykevich SA, Bell CM, Kaboli PJ, Auerbach AD, et al. Hospital readmission in general medicine patients: a prediction model. J Gen Intern Med. 2010;25(3):211–9. [PubMed: 20013068]

34. Vapnik VN. The nature of statistical learning. New York: Springer, 1995.

35. Daniel G Principles of artificial neural networks. Singapore: World Scientific Press, 2013.

36. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002;2(3):18–22.

37. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001:1189–232.

38. Ye J, Chow J-H, Chen J, Zheng Z. Stochastic gradient boosted distributed decision trees. Proceedings of the 18th ACM conference on Information and knowledge management; ACM, New York 2009, 2061–2064.

39. Daghistani TA, Elshawi R, Sakr S, Ahmed AM, Al-Thwayee A, Al-Mallah MH. Predictors of in-hospital length of stay among cardiac patients: A machine learning approach. Int J Cardiol. 2019;288:140–7. [PubMed: 30685103]

40. Ramkumar PN, Navarro SM, Haeberle HS, Karnuta JM, Mont MA, Iannotti JP, Patterson BM, Krebs VE. Development and validation of a machine learning algorithm after primary total hip arthroplasty: applications to length of stay and payment models. J Arthroplasty. 2019;34(4):632–7. [PubMed: 30665831]

41. Hopkins BS, Yamaguchi JT, Garcia R, Kesavabhotla K, Weiss H, Hsu WK, Smith ZA, Dahdaleh NS. Using machine learning to predict 30-day readmissions after posterior lumbar fusion: an NSQIP study involving 23,264 patients. J Neurosurg: Spine. 2019;1(aop):1–8.

42. Nakazawa KR, Cornwall JW, Rao A, Han DK, Ting W, Tadros RO, et al. Trends, Factors, And Disparities Associated With Length of Stay After Lower Extremity Bypass For Tissue Loss: Lessons Learned From NSQIP. J Vasc Surg. 2020;S0741–5214(20)31244–1.

43. Pagedar NA, Davis AB, Sperry SM, Charlton ME, Lynch CF. Population analysis of socioeconomic status and otolaryngologist distribution on head and neck cancer outcomes. Head Neck. 2019;41(4):1046–52. [PubMed: 30549368]

44. Amini A, Verma V, Li R, Vora N, Kang R, Gernon TJ, et al. Factors predicting for patient refusal of head and neck cancer therapy. Head Neck. 2020;42(1):33–42. [PubMed: 31584746]

45. Stubbs VC, Rajasekaran K, Cannady SB, Newman JG, Ibrahim SA, Brant JA. Social determinants of health and survivorship in parotid cancer: An analysis of the National Cancer Database. Am J Otolaryngol. 2020;41(1):102307. [PubMed: 31732319]

46. Bates JE, Hitchcock KE, Mendenhall WM, Dziegielewski PT, Amdur RJ. Comparing national practice versus standard guidelines for the use of adjuvant treatment following robotic surgery for oropharyngeal squamous cell carcinoma. Head Neck. 2020;42(9):2602–6. [PubMed: 32476219]

47. Osborn VW, Givi B, Rineer J, Roden D, Sheth N, Lederman A, et al. Patterns of care and outcomes of adjuvant therapy for high-risk head and neck cancer after surgery. Head Neck. 2018;40(6):1254–62. [PubMed: 29451961]

48. Chen MM, Roman SA, Sosa JA, Judson BL. Histologic grade as prognostic indicator for mucoepidermoid carcinoma: A population-level analysis of 2400 patients. Head Neck. 2014;36(2):158–63. [PubMed: 23765800]

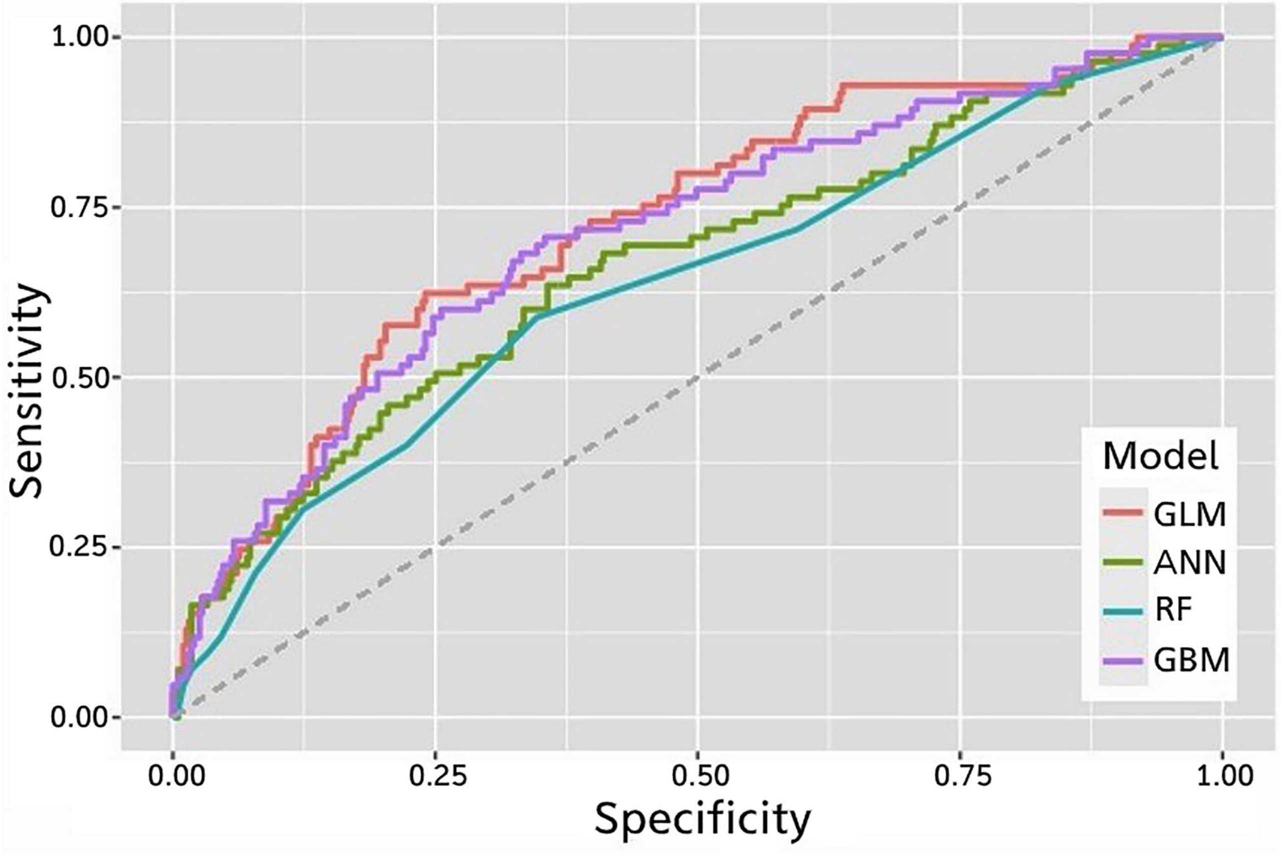

Figure 1.

Receiver operating characteristic of the GLM, ANN, RF, and GBM models that used pre-operative features to predict DNHF (AUC range=0.66–0.73)

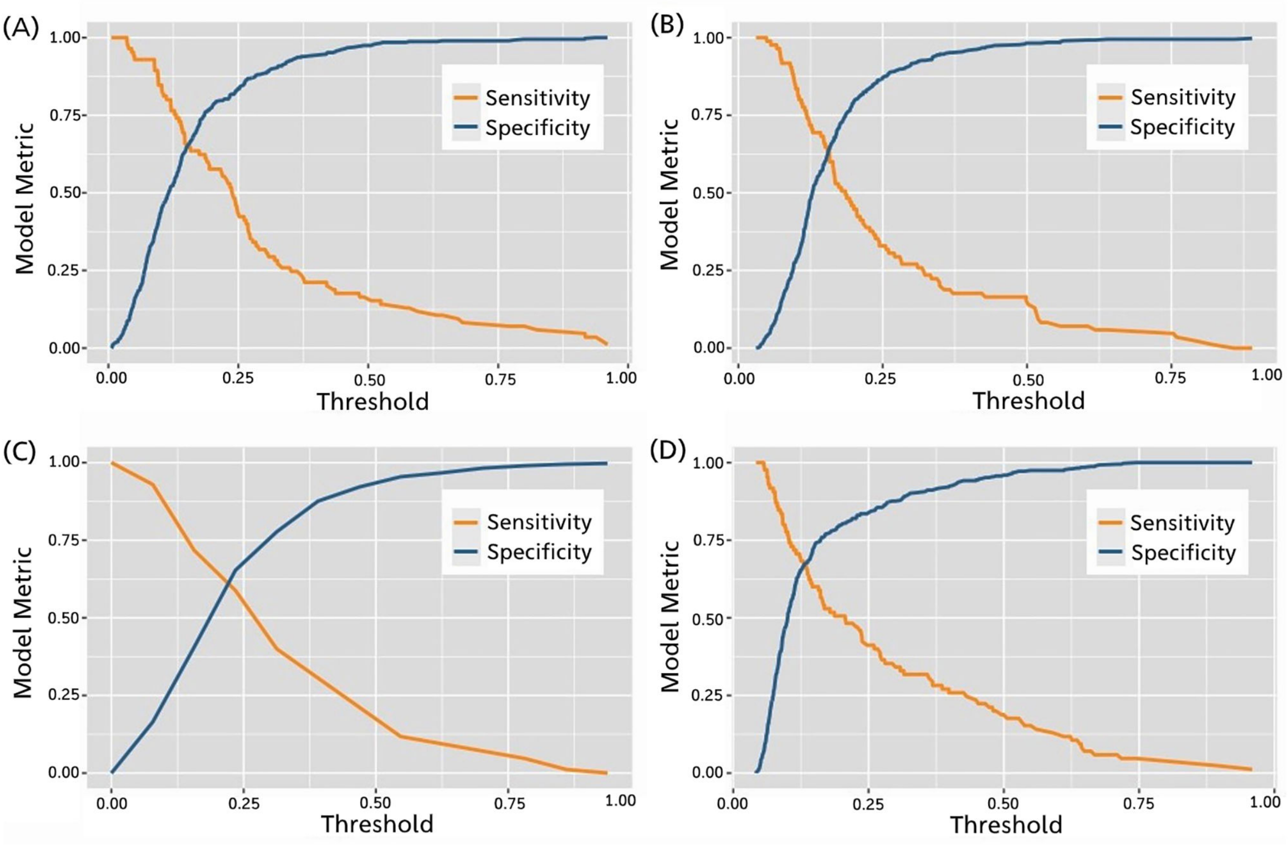

Figure 2.

The trade-off relationship between sensitivity and specificity of the constructed models, directly dependent on the designated classification threshold of a given model, are demonstrated for (A) GLM, (B) ANN, (C) RF, and (D) GBM models.

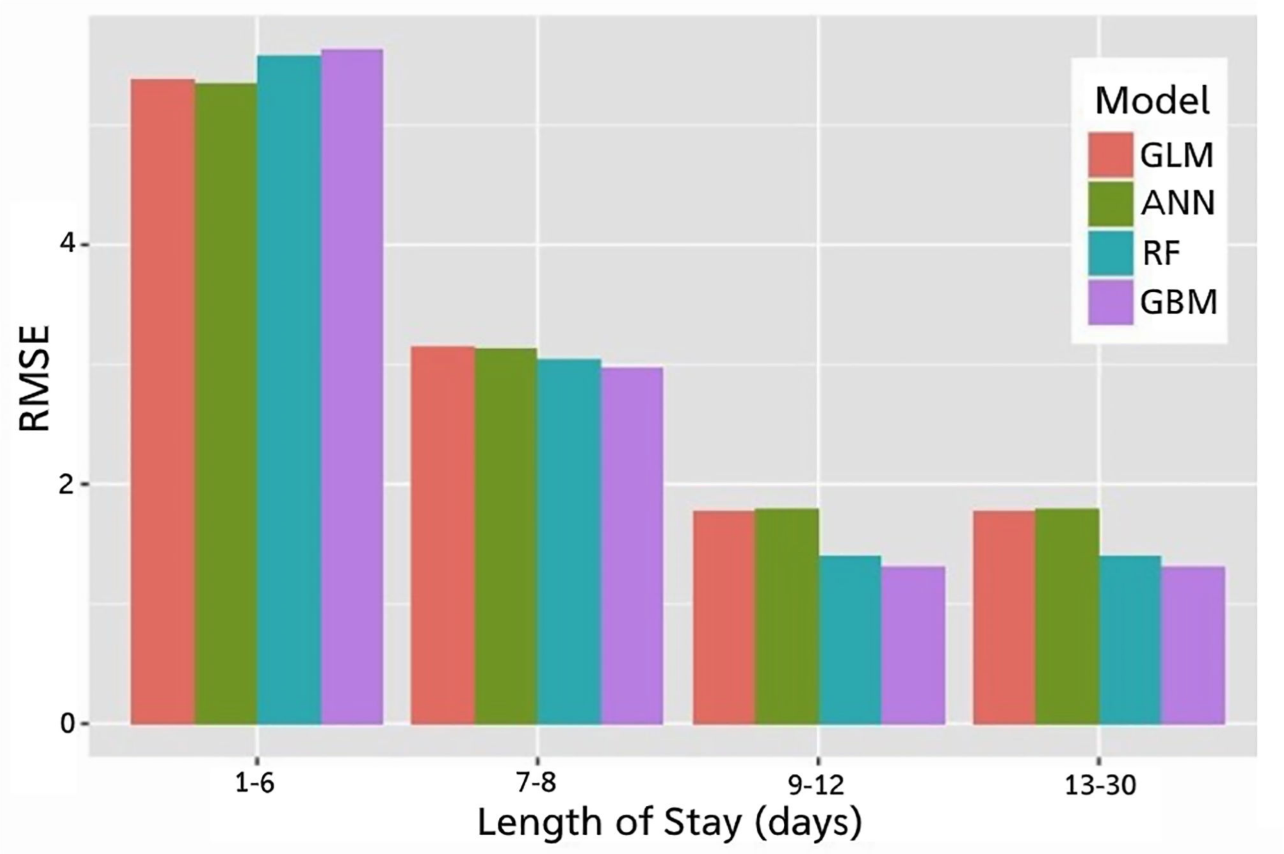

Figure 3.

Comparing RMSE of the regression models stratified based on LOS quartiles: 1–6 days (N=105), 7–8 days (N=154), 9–12 days (N=131), and 13–30 days (N=144). The stratification is according to the actual LOS, and the bars represent the root mean squared error of the respective predicted LOS.

Table 1.

Variables included in the machine learning models predicting DNHF, compared via univariate analysis (chi-square or independent *t*-test).

Pre-Operative Feature	**Discharge to Home (N=1978)**	**Discharge to Nonhome Facility (N=421)**	***P* Value**
Age (mean ± SD)	60.7 ± 11.4	67.7 ± 11.3	**<0.001**
BMI (mean ± SD)	25.6 ± 6.3	24.5 ± 5.8	**<0.001**
Gender: Female	536 (27.1)	126 (29.9)	0.236
Procedure: Laryngectomy	1108 (56.0)	242 (57.5)	0.575
Diabetes	265 (13.4)	66 (15.7)	0.217
Smoking	819 (41.4)	163 (38.7)	0.312
Dyspnea	250 (12.6)	98 (23.3)	**<0.001**
Functional status: Dependent	36 (1.8)	46 (10.9)	**<0.001**
Ventilator dependent	9 (0.5)	6 (1.4)	**0.022**
History of COPD	213 (10.8)	81 (19.2)	**<0.001**
History of CHF	18 (0.9)	11 (2.6)	**0.004**
History of wound infection	52 (2.6)	21 (5.0)	**0.010**
Hypertension	925 (46.8)	241 (57.2)	**<0.001**
ASA class: High	1651 (83.5)	389 (92.4)	**<0.001**
Chronic steroid use	75 (3.8)	30 (7.1)	**0.002**
Emergency surgery	15 (0.8)	5 (1.2)	0.378
Elective surgery	1836 (92.8)	352 (83.6)	**<0.001**
Systemic sepsis	31 (1.6)	19 (4.5)	**<0.001**
Pre-operative transfusion	11 (0.6)	6 (1.4)	0.053
Pre-operative sodium (mean ± SD)	138.4 ± 3.7	137.7 ± 4.0	**0.001**
Pre-operative BUN (mean ± SD)	15.9 ± 9.0	17.6 ± 9.9	**0.002**
Pre-operative creatinine (mean ± SD)	0.90 ± 0.53	0.86 ± 0.36	0.114
Pre-operative WBC (mean ± SD)	8.0 ± 3.0	8.5 ± 3.5	**0.003**
Pre-operative HCT (mean ± SD)	38.7 ± 5.2	36.8 ± 5.3	**<0.001**
Pre-operative platelets (mean ± SD)	265.7 ± 96.6	279.7 ± 98.4	**0.008**

Values in parenthesis are percentages.

DNHF: discharge to nonhome facility; SD: standard deviation; BMI: body mass index; COPD: chronic obstructive pulmonary disease; CHF: congestive heart failure; ASA: American Society of Anesthesiologists; BUN: blood urea nitrogen; WBC: white blood cells; HCT: hematocrit.

Table 2.

Performance of different classification machine learning models predictive of DNHF.

Model	Accuracy	Sensitivity	Specificity	AUC-ROC	PPV	NPV
GLM	0.7542	0.5294	0.8025	0.7253	0.3659	0.8880
ANN	0.7316	0.4200	0.7986	0.6719	0.3419	0.8735
RF	0.7603	0.3894	0.8401	0.6639	0.3451	0.8648
GBM	0.7623	0.4712	0.8249	0.7281	0.3666	0.8788

DNHF: discharge to nonhome facility; GLM: generalized linear model; ANN: artificial neural network; RF: random forest; GBM: gradient boosting machine; AUC-ROC: area under the curve of receiver operating characteristic; PPV: positive predictive value; NPV: negative predictive value.

Table 3.

Performance of different regression machine learning models predictive of LOS in days.

Regression Model	MAE	MSE	RMSE	RMSLE
GLM	3.9559	26.5832	5.1559	0.4545
ANN	3.9456	26.5878	5.1563	0.4536
RF	3.9770	26.4624	5.1442	0.4590
GBM	3.9783	26.6121	5.1587	0.4610

LOS: length of stay; GLM: generalized linear model; ANN: artificial neural network; RF: random forest; GBM: gradient boosting machine; MAE: mean absolute error; MSE: mean squared error; RMSE: root mean squared error; RMSLE: root mean square logarithmic error.

Table 4.

Logistic regression weights of model features with the most contribution.

DNHF Model		LOS Model	
Variable	**Odds ratio**	**Variable**	**Odds ratio**
Functional status	4.66	Pre-operative transfusion	9.86
Pre-operative transfusion	3.01	Elective surgery	6.45
Ventilator dependent	2.29	Procedure type: resection	6.04
History of CHF	2.06	Disseminated cancer	3.63
History of wound infection	2.04	History of CHF	3.65
Pre-operative creatinine	1.98	Functional status	2.71
Elective surgery	1.74	Chronic smoking	2.30
Procedure type: resection	1.63	Systemic sepsis	1.74
ASA class	1.40	Diabetes	1.62
Dyspnea	1.40	Race	1.57
Chronic steroid use	1.35	ASA class	1.26
COPD	1.34	History of wound infection	1.20
Race	1.22	Pre-operative sodium	1.13

DNHF: discharge to nonhome facility; LOS: length of stay; CHF: congestive heart failure; ASA: American Society of Anesthesiologists; COPD: chronic obstructive pulmonary disease.

Odds ratios signify the contribution of these variables and not necessarily the direction of the association.