A Machine Learning Algorithm Using Clinical Perioperative Features Performed with Very High Accuracy in the Prediction of In-Hospital Mortality

Helena Linnen & Kahye Song

Received August 26, 2024
Accepted January 14, 2025
Electronic access January 31, 2025

Background/Objective: Medical researchers have built predictive models to predict patient mortality after surgery. By identifying patient risk of death, clinicians are better prepared to attempt treatments to save the patient. We pursued three specific aims: First, to quantify the predictive capabilities of XGBoost, a machine learning algorithm, in a surgical population of patients. Second, to determine whether the model offers clinical utility. Third, to compare our findings with existing results.
Methods: Our research sample included 6,314 adult patients with non-cardiac surgeries in a South Korean tertiary university hospital, extracted electronically from the VitalDB database. We trained and tested a model predicting in-hospital mortality after surgery, which included 30 clinical feature variables. We randomized patients into two model partitions, and summarized model performance. Our model utilized XGBoost, a machine learning library with superior classification performance.
Results: We observed a total of 52 in-hospital deaths. Among clinically relevant features were age, sex, hypertension, diabetes, abnormal electrocardiogram or pulmonary function test, abnormal liver and arterial blood gas results. Model accuracy for the XGBoost model was 99%, the F1-score was 0.91, recall was 0.91, precision was 0.91, and the AUROC was 74%. The confusion matrix demonstrated good clinical utility of the model in that it correctly classified both survivors and decedents well.
Conclusion: Our findings agree with previous research in the field of post-surgical mortality prediction using machine learning. Clinicians may benefit from the risk identified to make informed treatment decisions with their patients and families. Future research should aim to pool international datasets to overcome sample size restrictions and sample homogeneity.

Keywords: Surgery, In-hospital death, predictive analytics, machine learning, clinical decision support

Background and Significance

For physicians and hospitalized patients, safety is the number one priority. However, every year in the United States, over 250,000 patients die unexpectedly from iatrogenic events which includes negligence in regard to equipment, personnel, procedures and management [1]. Researchers have built a variety of predictive models in the hopes of predicting patient mortality preemptively [2–5]. By identifying patient risk of death, surgeons are better prepared to attempt treatments to save the patient [2–5]. A smaller subset of surgical patients offers a unique population that is by definition quite different from other hospitalized patients. Predicting patient mortality among surgical patients has shown promise, but the majority of predictive models were derived in single centers [3–5]. Therefore, due to model bias, prediction of mortality is not necessarily transferable to other patient populations in other hospitals [6].

While models exist that can predict mortality after surgery [2–5], none can be universally used across country (or even state) borders. Few studies have used large sample sizes [3] and even fewer have developed predictive mortality models for patients after surgery. Large sample sizes are important because they provide more accurate information about a given population [7]. Given some deaths are considered avoidable [7], it is important to investigate whether preoperative clinical data could be useful in the prediction of patient mortality. The awareness that a patient may be at higher risk of death could inform preemptive clinical support measures, including planning for post-operative admission to intensive care, or inform the urgency/need for a given surgery, especially those that are purely elective.

This research study was hypothesis generating. No theory of clinical deterioration risk and clinical decision support exists to our knowledge. The field of clinical predictive modeling is relatively novel with a large increase in publication in the past decade. Our aim was to explore potential risk-factors in regard to preoperative mortality prediction and to compare our findings with the field. We did not employ inferential statistics to test a hypothesis against an existing theory or framework. Rather, we aimed to evaluate predictive model features, feature importance, model interpretability, and predictive power in the prediction of

postsurgical mortality among hospitalized patients.

Problem Statement

The majority of the evidence comes from single-center studies without secondary validation in another research population. For example, Protopapa et al. (2014)[4] described the development and validation of a risk tool for post-surgical mortality in Australia. The SORT model was externally validated[8] but was also found to not be fully representative of the population in which it was validated (Australia). Due to the apparent model bias in the original model, it would be important to further examine additional populations in other parts of the world. For example, Huber et al. (2024)[9] described their predictive model based on a South Korean patient population. Given that their findings are novel, the field would benefit from secondary validation.

Our study fills a significant gap: Given how few studies exist that use machine learning in the prediction of in-hospital mortality after surgery[2–5] and the fast advancement of machine learning algorithms, a considerable evidence gap exists. While Huber et al. (2024)[9] recently published their predictive model, nevertheless it is prudent to validate the predictive features and properties within the same population using an equally or more sophisticated machine learning method.

Objectives

We pursued three specific aims: First, to quantify the predictive capabilities of XGBoost, an advanced machine learning algorithm, in a surgical population of patients in South Korea. Second, to determine whether the model offers new clinical utility. Third, to compare our findings with existing results.

Our study did not require IRB approval because all patient data were publically available and fully de-identified[7]. For the purposes of this research study, it was not necessary to acquire identifiable data or conduct recruitment procedures. The IRB at Seoul National University Hospital in Seoul, South Korea approved the acquisition and release of the data in its original form.

Methods

We included patients whose pre-operative data were recorded in the publicly available vitalDB[10] dataset in South Korea. To minimize bias we excluded patients younger than 18 years of age because they present with different physiological features than adult patients. All patients underwent surgery and we did not limit the dataset to specific surgery types. Equally, we included all surgery durations.

In this data-only study, we developed a predictive model to identify in-hospital mortality after surgery using a machine learning classification model. We used 30 clinical feature variables, randomized patients into two model partitions, and summarized model performance.

Research Design

This study used an observational cohort design. We extracted data from VitalDB11, an open-source clinical data repository of de-identified perioperative patient data at Seoul National University Hospital, Seoul, Republic of Korea. We included patients with non-cardiac surgeries between August 2016 to June 2017 (10 months). In total, the dataset includes 6,388 cases of which we retained 6,314 cases with adults. We obtained data from non-cardiac surgery patients (general, thoracic, urologic, and gynecologic) who underwent routine or emergency surgery at Seoul National University Hospital, Seoul, Republic of Korea. The dataset contains a total of 557,622 (average 87, range 16-136) data tracks from 6,388 cases. We extracted these data tracks in csv (comma separated value) format.

Variables and Measurements

To measure patient mortality after surgery, we selected hospital mortality ('death_inhosp'), a binary flag variable that indicates whether a patient died during the hospitalization following surgery. The variable is all-cause meaning that the dataset did not differentiate potential different causes of death. For purposes of this research study, and in the interest of parsimony, we elected to accept this definition.

We included the following predictive feature variables: Age, body mass index (weight in kg divided by height in m^2), sex, operation type, diagnosis, operation name, flags for preoperative hypertension and diabetes mellitus, preoperative electrocardiogram diagnosis (e.g., normal sinus rhythm), and preoperative pulmonary function test. Preoperative lab work (serum): hemoglobin, platelet count, prothrombin time, activated partial thromboplastin time, sodium, potassium, glucose, albumin, aspartate transferase, alanine transferase, blood urea nitrogen, and creatinine. Preoperative arterial blood gas results: pH, bicarbonate, base excess, partial pressure of O2, partial pressure of CO2, and arterial oxygen saturation[10].

After we extracted data from VitalDB, and excluded non-adult patients, we used Python[11] within the GoogleCollab[12] integrated development environment to develop the predictive model. We used the following library packages: Pandas[13], MatPlotLib[14], Scikit-Learn[15], NumPy[16], and requests[17].

To derive and validate the predictive model, we employed eXtreme Gradient Boosting (XGBoost)[18], a machine learning library that uses a scalable, distributed gradient-boosted decision tree[19]. A decision tree uses probabilistic branching to determine the likelihood of a given feature variable on the outcome. It is called a "tree" because of its branchlike logical structure where one feature can be hierarchically distributed above or below

another feature. XGBoost uses an algorithm similar to a Random Forest[19] structure, which replicates k numbers of decision trees and iteratively refines the prediction by taking multiple weak models and training each model using the error term of the previous model (boosting). Through numerous cycles, this technique reduces a model's error term to the largest degree currently known[20]. Specifically, XGBoost builds consecutive trees with each step reducing the error of the iteration before it. We selected XGBoost because of these properties and because it fully met the needs of the project.

After randomization, we split the dataset into a training and testing set (70% and 30%, respectively). We then trained the model, and gathered model performance metrics including the Area Under the Receiver Operating Characteristic Curve (AUROC) and model accuracy. Finally, we plotted the confusion matrix and AUROC graphically. The AUROC compares the model's false positive rate (x-axis) with its false negative rate (y-axis). It measures how well predictions are ranked and the quality of the model's predictions[21]. Accuracy is a model performance metric that measures the number of correct predictions as a percentage of the total number of predictions[22].

Results

Table 1 describes the cohort characteristics. We observed substantial differences between descendants and survivors. Among the most notable were the deltas for age, sex, past hypertension, diabetes mellitus, abnormal electrocardiogram, abnormal pulmonary function test, aspartate transferase, alanine transferase, arterial base excess, and arterial partial pressure of O2. These results make intuitive sense because one would expect older patients or those with pre-existing heart conditions and other comorbidities to fare worse after surgery. Among laboratory values, our findings suggest that patients with elevated liver enzymes are at higher risk of death as well (this result may point to high alcohol use among men in this cohort). Similarly, patients with abnormal pulmonary function tests are likely current or former smokers, although the dataset did not offer a definitive answer. Furthermore, they enhance our understanding as to which clinical factors are most predictive. For example, we observed that Aspartate Transferase (a liver enzyme) is a much more influential factor than age. This awareness enables healthcare providers to better understand how different risk factors affect in-hospital mortality after surgery.

Of the 6,314 patients with non-cardiac surgeries, we observed a total of 52 in-hospital deaths. Therefore, the pretest probability of death in this patient cohort was 0.82%. After randomization and splitting of the data, we observed a good balance of the outcome variable in both partitions (64% of observed deaths in the 70% testing partition, which equals an 8.6% relative between-partition imbalance). Due to the rarity of the outcome, a balanced representation of death in both partitions is crucial to develop a sound model. The XGBoost model resulted in the following performance: Model accuracy was 99% indicating very good ability to determine the total number of predictions correctly. This finding suggests that the model makes a significant amount of its decisions correctly when assessing mortality. The AUROC was 74% indicating moderate ability to discriminate the false positives against the false negatives (see Figure 1). It is likely a function of the very low pretest probability and small number of outcomes, that the AUROC did not score higher. The confusion matrix for the testing partition (see Figure 2) demonstrates that the model could predict in-hospital death in 90% of patients who died, those being only 8 in 1,000 patients, and it could predict survival nearly perfectly. The F1-score was 0.91, recall was 0.91, precision was 0.91 (see Table 2)

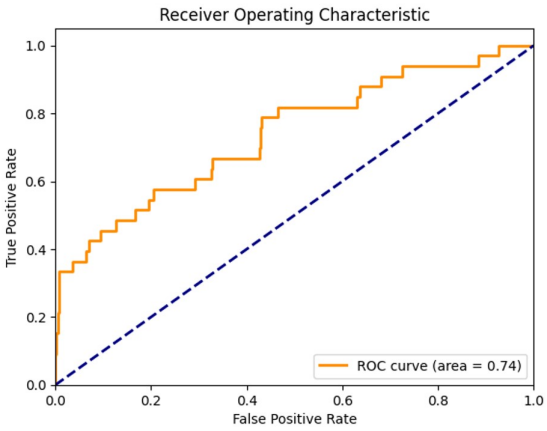

Fig. 1 XGBoost Area Under the Receiver Operating Curve for the Prediction of Postoperative Hospital Mortality.

Notes. *AUROC visualizes the increase of a model's true positive rate (those who died and were predicted to die) against the increase in the model's false positive rate (those who lived but were predicted to die). The graph shows the model predicts substantially better than chance (dotted blue line).*

Our results suggest that, in the research population of surgical patients in a university hospital in a South Korea metropolitan area, the model could have real-world application and benefit. For surgeons and other treating providers, knowing with reasonable certainty that a patient is likely not to survive after surgery, may be an important datapoint. For example, clinicians could weigh the surgical benefits and risks in a better light and inform decision making between the patient, family, and clinicians. That said, physicians cannot over-rely on any predictive model due to the inherent limitations. Linnen et al.[6] cautioned that predictive models should be used as additional diagnostic tools, but not be used in place of a physician's judgment.

Shapley Additive exPlanations (SHAP) is a computer programming approach to help in the determining feature importance in a machine learning model. As Figure 3 shows, certain

Table 1 Cohort Characteristics

Variables	Entire Cohort	Decedents	Survivors	Delta1 (Decedents - Survivors)
N (%)2	6,314 (100)	52 (0.8)	6,262 (99.2)	n/a
Age, mean (Std.Dev3)	57.8 (14.2)	61.2 (15.6)	57.8 (14.2)	(3.4)
Sex, n (%)	M: 3,205 (50.8) F: 3,109 (49.2)	M: 37 (71.2) F: 15 (28.2)	M: 3,168 (50.6) F: 3,094 (49.4)	M: (20.6) F: (-21.2)
Past Hypertension (flag) (%)	1,964 (31.1)	19 (36.5)	1,945 (31)	(5.5)
Diabetes Mellitus (%)	661 (10.5)	3 (5.7)	658 (10.5)	(-4.8)
Abnormal Electrocardiogram (%)	81 (1.3)	9 (17.3)	72 (1.5)	(15.8)
Abnormal Pulmonary Function Test (%)	1,038 (16.4)	13 (25)	1,025 (16.3)	(8.7)
Aspartate Transferase, mean (Std.Dev)	31.1 (147.4)	125.5 (294.5)	30.4 (145.5)	(149)
Alanine Transferase, mean (Std.Dev)	28.7 (94.5)	99.0 (203.1)	28.1 (92.9)	(110.2)
Base Excess, mean (Std.Dev)	-0.3 (3)	24.0 (1.7)	-0.3 (3.1)	(-1.4)
Partial Pressure of O2, mean (Std.Dev)	104.4 (44.1)	91.9 (37.2)	104.5 (44.2)	(-7)

1. The difference in either the percentage or the mean result between those who died and those who lived to discharge
2. Percentages are displayed in parentheses and calculated for each column based on the denominator described (all, decedents, survivors).
3. Standard Deviation
4. Body Mass Index

Table 2 XGBoost Confusion Matrix Results for the Prediction of Postoperative Hospital Mortality

	Predicted Correctly[1]	Predicted Incorrectly	Model Performance
Decedent[2]	30	3	Precision: 0.91 Recall: 0.91 F1: 0.91
Survivor	4384	3	

[1] A correct prediction indicates that the model categorized a given patient into the right actually observed outcome group.
[2] Those who died in-hospital.

features (risk factors) had a prominent influence on predicting patient mortality. We confirmed that the SHAP results corresponded with observed delta values from Table 1. In combination, and with the support of a clinically trained subject matter expert, we concluded that the SHAP values are both plausible and clinically meaningful. For example, abnormal electrocardiogram is shown to be a feature in its delta in Table 1 and in the SHAP waterfall diagram. Its delta is 15.8% and value of magnitude is 19.6 in the waterfall diagram. What this means, is that decedents are 15.8% to have electrocardiograms than survivors. This is validated by the SHAP diagram because it also shows the strong negative effect.

Discussion

Our results agree with previous work in the field of mortality prediction among surgical populations. Glance et al.[2] developed a 30-day mortality risk index prediction model for noncardiac surgery using data from the American College of Surgeons National Surgical Quality Improvement Program database. The reported C statistic was 0.897 (AUROC equivalent). It is likely that their C-statistic was higher than our model because they had more than a 4.7 times larger dataset (nearly 300,000 patients in total). Their population was also more diverse, given the American College of Surgeons National Surgical Quality Improvement Program database spans 643 hospitals across the United States[23]. Le Manach et al.[3] developed a model (Preoperative Score to Predict Postoperative Mortality) that predicted

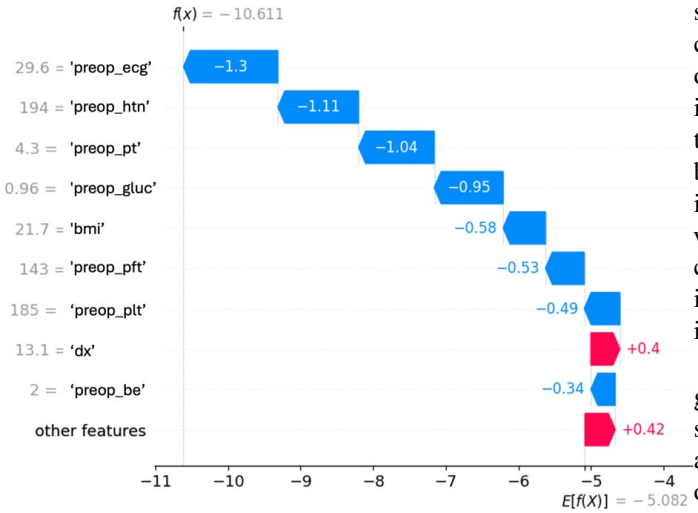

Fig. 2 Waterfall Diagram of Clinical Feature Importance in a Model Predicting Postoperative Hospital Mortality.

Notes.

1. ECG = electrocardiogram; HTN = hypertension; PT = prothrombin time; GLUC = glucose; BMI = body mass index; PFT = pulmonary function test; PLT = platelet count; DX = diagnosis; BE = arterial blood gas base excess
2. The diagram depicts feature importance of the XGBoost model ranked from highest to lowest.

in-hospital mortality using data from French hospitals. The reported C statistic was 0.929. Again, the dataset used was 87 times larger and was widely spread across France compared to VitalDB's single center at the Seoul National University Hospital. Protopapa et al.[4] developed a model (Surgical Outcome Risk Tool) that predicted 30 day mortality after non-cardiac surgery using National Confidential Enquiry into Patient Outcome and Death Knowing the Risk study. The reported AUROC was 0.91. Some patients do not choose to stay in the hospital after surgery as they can return home or go to rehabilitation. By broadening the requirements needed to qualify for a death, it is likely to have greatly contributed to such a high AUROC. Finally, Campbell et al.[5] developed a model (New Zealand Risk model) that predicted 30 day, 1 year, and 2 year mortality after surgery using the New Zealand National Minimum Data Set. The reported AUROC was 0.921 for 1-month mortality. Similarly to the VitalDB dataset, Campbell et al.'s model was also trained on hospital data in a small number of hospitals in New Zealand. However, the AUROC is for 1 month mortality, which is a much greater timeframe and does not require the patient to die in-hospital. This AUROC is likely higher than our model because of these two favorable changes.

The field of mortality risk prediction for surgical patients using AI tools is relatively new. In our literature review the oldest publication dated back to 2012[2]. The value of our research study is twofold: First our findings enhance and validate the clinical understanding of risk factors and confirm results previously published. Second, our study is clinically useful, because it demonstrates how predictive analytics can assist surgeons in their risk assessment and discussion with patients and family before surgery. Among the most important aspects of our study is the very high model accuracy. The performance metrics revealed that even with a limited sample, but a set of well curated clinical features, it is possible to correctly predict mortality risk in a majority of patients. This validation is significant because it corroborates findings of existing published research.

Furthermore, while this model is not able to replace a surgeon's decision making, it is able to serve as a clinical decision support tool. The model can inform, which patients may be at higher risk of death before surgery (for example, based on our model, an overweight 82-year-old male who frequently consumes alcohol and tobacco). In South Korea, a male has a life expectancy of 79.9 in 2022[24]. As can be seen by Table 1. And Figure 3, obesity, low pulmonary function, significantly higher Aspartate and Alanine Transferase are much more likely to occur in a decedents than survivors. Such information could assist the surgeon or physician deciding with the patient about the risks and benefits of a given surgery.

Our research study successfully met its three specific aims. We quantified the predictive capabilities of XGBoost, determined its clinical utility, and compared our findings with notable studies in the field. Based on our findings, even with limited data, machine learning models are impressively capable in supporting clinical decisions. Other studies confirmed our methodology and findings with similar AUROC values.

Limitations

We would like to point to four limitations in this research study. First, our findings were derived from a specific single-center research population in South Korea. Our results may have limited generalizability because the relative homogeneity of patient characteristics, single-center clinical services and procedures. These factors may limit transferability of results to other clinical settings and/or countries. Even though a South Korea metropolitan university hospital may produce different cohort characteristics, our results are very plausible given its agreement with other international work in the field.

Second, the available sample size was relatively small and the outcome variable was very rare. Fortunately, the dataset was of very high quality which allowed us to investigate a very substantial number of clinical features. Even though in-hospital mortality was a rare outcome, our sampling procedures confirmed a balanced allocation of outcomes in both the training and testing partitions. Ultimately, the model performance results confirmed very good predictive capabilities despite this limitation.

Third, the outcome variable of in-hospital mortality may exclude patients in whom death occurred after they were discharged from the hospital. It is known6 that 30/60/90-day mortality would offer a more complete understanding of the outcome. However, the strong accuracy of our model suggests that it could equally well predict other mortality definitions.

Fourth, as a University hospital, the types of surgeries performed may be more expansive than those in community hospitals (e.g., transplant surgery). The VitalDB dataset, on which we build our model, may also bias results for rare surgery types. Only 236 different types of operations and procedures are included in this dataset, with some of them only occurring 1- 4 times in the whole dataset. Some operations and procedures are also more prevalent occurring more than 50 times in the dataset.

Conclusion

Future research should examine the development of an international data repository where multiple research centers can pool data. Such aggregate data could include preoperative or postoperative data, or even data beyond a patient's individual hospital stay. We also recommend examining time-based events such as a flagging system for data values that are outside of a given temporal range, and data streaming from medical devices during surgery. Finally, examining the outcome definition through an extended lens (30/60/90-day mortality) may glean a more complete picture of true patient death or survival following surgery. Patient death is arguably the most important aspect of clinical decision-making for healthcare providers and their patients. Our model is able to discriminate patients at high risk of post-operative death, while its application and transferability is limited to like tertiary medical centers with similar patient characteristics. While predictive models can serve as a decision support system, they cannot be used to replace of physicians' authority and judgment. To overcome single center sample size restrictions, it is important to further invest into creating internationally pooled datasets and more holistic models to predict mortality.

Acknowledgment

The authors would like to thank Daniel Linnen, PhD for his support throughout the journey of developing this paper and for sharing his research expertise and clinical subject matter expertise in interpreting model results.

References

1 M. A. Makary and M. Daniel, *Medical error-the third leading cause of death in the US*, 2016.

2 L. G. Glance *et al.*, *The Surgical Mortality Probability Model: derivation and validation of a simple risk prediction rule for noncardiac surgery*, 2012.

3 Y. Le Manach *et al.*, *Preoperative Score to Predict Postoperative Mortality (POSPOM): Derivation and Validation*, 2016.

4 K. L. Protopapa *et al.*, *Development and validation of the Surgical Outcome Risk Tool (SORT)*, 2014.

5 D. Campbell *et al.*, *National risk prediction model for perioperative mortality in non-cardiac surgery*, 2019.

6 D. T. Linnen *et al.*, *Statistical Modeling and Aggregate-Weighted Scoring Systems in Prediction of Mortality and ICU Transfer: A Systematic Review*, 2019.

7 T. G. Weiser and A. Gawande, *Excess Surgical Mortality: Strategies for Improving Quality of Care*, 2015, https://www.ncbi.nlm.nih.gov/books/NBK333498/.

8 J. R. Reilly *et al.*, *External validation of a surgical mortality risk prediction model for inpatient noncardiac surgery in an Australian private health insurance dataset*, 2022.

9 M. Huber *et al.*, *Decision Curve Analysis of In-Hospital Mortality Prediction Models: The Relative Value of Pre- and Intraoperative Data For Decision-Making*, 2024.

10 H. Lee, Y. Park, S. Yoon, S. Yang, D. Park and C. Jung, *VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients*, 2022.

11 Python Software Foundation, *Python Language Reference, version 2.7*, 2024, http://www.python.org.

12 Google, *Google Colaboratory*, 2024, https://colab.research.google.com/.

13 The pandas development team, *pandas*, 2023, https://doi.org/10.5281/zenodo.3509134.

14 J. D. Hunter, *Matplotlib: A 2D Graphics Environment*, 2007.

15 F. Pedregosa *et al.*, *Scikit-learn: Machine Learning in Python*, 2011.

16 C. R. Harris *et al.*, *Array Programming with NumPy*, 2020.

17 K. Reitz, *Requests: HTTP for Humans*, 2023, https://requests.readthedocs.io.

18 T. Chen and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, 2016.

19 L. Breiman, *Random Forests*, 2001.

20 NVIDIA, *What Is XGBoost?*, 2024, https://www.nvidia.com/en-us/glossary/xgboost/.

21 Google, *Classification: ROC Curve and AUC*, 2022, https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

22 Google, *Classification: Accuracy*, 2022, https://developers.google.com/machine-learning/crash-course/classification/accuracy.

23 ACS, *Hospital and Facilities*, 2024, https://www.facs.org/hospital-and-facilities/?searchTerm=&institution=NsqipHospital&address=&sort=a-z&page=1.

24 L. Yoon, *Life expectancy of men at birth in South Korea from 1970 to 2022*, 2024, https://www.statista.com/statistics/1040739/south-korea-life-expectancy-of-men/#statisticContainer.