

Predicting the incidence of postoperative nausea and vomiting

Ken Marciel, Project in Statistics (STAT 685), Texas A&M University
Spring 2022

The incidence of post-operative nausea and vomiting (PONV) is generally in the range of 20-40% (Apfel *et al*). This condition has negative effects on the health and well-being of patients and is financially costly to healthcare providers. The tradeoff is that preventive therapy has negative side effects and financial costs. So, the challenge is to develop a scoring system that most accurately recommends prophylaxis for the patients at high risk of PONV. In other words, a predictive model that is neither too conservative nor too liberal in determining which patients should be prescribed prophylaxis. This balancing act has been described in the medical literature as the *prevent-or-cure dilemma*.

There are several well-documented models for predicting PONV, to help guide prudent administration of anti-emetic prophylaxis. These models are typically developed using logistic regression and stepwise backward elimination for variable selection. The most common measures of validity are discrimination and calibration. The most common measure of discriminating power is AUC, the area under the receiver operating characteristic curve (ROC). Calibration is most commonly assessed using the slope and squared correlation (R^2) for the line in a calibration plot. In the literature, AUC values range from 0.61 to 0.785 (Apfel *et al*, Sinclair *et al*), calibration slopes range from 0.3 to 1.71 (Apfel *et al*, Eberhart *et al*), and squared correlation ranges from 0.763 to 0.99 (Apfel *et al*).

This investigation analyzed a data set of 461 patients from anesthesiologist Jelena Velickovic, MD, in Belgrade, Serbia. The purpose was to develop a predictive model for PONV with performance comparable to or better than models previously published in the medical literature.

Methods

Software

Data analysis was performed using the R statistical computing software through the R Markdown interface, with the following additional packages installed to R.

```
library(dplyr) # rename variables
library(alr4) # marginal model plots
library(leaps) # regression subset plots
library(car) # regression subset plots
library(rms) # logistic regression
library(pROC) # ROC curve
library(caret) # data splitting, resampling
```

Data set

The raw data set has 916 rows and 26 columns. None of the rows have missing values. After removing the 93 duplicates, the cleaned data set has 823 rows and 26 columns.