Can we forecast lung cancer data accurately?

A Smith

October 2020

Executive summary

Using patient referral/test data for lung cancer from hospitals in the Cwm Taf Morgannwg University Health Board (CTMUHB) in South Wales, this project examines the extent to which we can accurately forecast future hospital referrals/tests using statistical analysis methods. Wales lags other countries for lung cancer outcomes, having the second-worst survival rates for the disease in Europe. The CTMUHB region has the highest occurrences of the disease per capita within Wales. As a result of this poor survival rate, NHS Wales has introduced the National Optimal Lung Cancer Pathway to expediate a patient's journey through the healthcare system, ensuring early diagnosis/treatment and improved outcomes. To ensure effective planning and improve efficiency, it is important to offer healthcare providers some estimate of the demand they may expect on services.

Our project stems from data of the details of patient interaction with the lung cancer pathway which contain the number of referrals/tests requested per day/week/month across the CTMUHB. We transform the data into a series of datapoints ordered by time, a *time series,* and perform rudimentary analysis. We also segment the data by hospital and test type to examine these categories independently. This initial analysis concludes that the Prince Charles and Royal Glamorgan hospitals receive the bulk of referrals, and that CT and PET tests are the most common tests undertaken. We create boxplots to determine if we can detect patterns in the frequency of tests/referrals by day/month. We found that the numbers of tests/referrals are largest during the working week, and that referrals and tests are more prevalent in spring/summer than winter. We continue our analysis by discussing *moving averages*, which smooth the data and provide insight into underlying patterns. Next, we perform statistical tests to determine whether the data is stationary (does not depend on time), or nonstationary (displays *seasonality*, a regular and predictable change, or a long-term *trend*). We find statistical evidence of nonstationarity on the monthly *referrals* and weekly *tests*. Finally, we decompose each series to examine the seasonality, trend, and *random error*.

We partition each series into a training set (to train each model) and a test set (to test its accuracy). We apply statistical forecasting methods to predict future values past the end of the existing series. The statistical methods used increased in complexity and are listed in order of complexity below:

1. ***Naïve methods***: which predict that last observed value will continue unchanged.

2. ***Extrapolation models:*** which apply a mathematical function to all preceding data points to forecast future datapoints and trends.
3. ***Causal models****:* which assume a linear relationship (line of best fit).
4. ***ARIMA models****:* which include parameters for seasonality and trend.
5. ***Singular spectrum analysis:*** which deconstructs the series into components to reduce *noise*.
6. ***Artificial neural network (ANN):*** a method inspired by the workings of the brain.

The accuracy of the forecasting models is assessed using several different *error statistics*, each provides different information about how well our forecasting models fit the data. We fit each model to each series and calculate the forecasts made and the error statistics produced. After analysis of each method, we endorse one method for each time series, based on its error statistics and a plot of its efficacy. We determine that ANNs are the method most often endorsed, though they may be unsuitable for linearly structured data. We find that a type of extrapolation model, *single exponential smoothing* (SES), is favourable against ANNs for non-noisy datasets. ANNs were effective on the noisy *daily* data, however other methods may be preferred for less variable *monthly* data.

Table 0.1: frequency table- most preferred forecast method.

Forecasting method	Frequency
Artificial neural network	6
Single exponential smoothing	3
Holt-Winters (extrapolation model)	2
Simple linear regression (causal)	2
Singular spectrum analysis	2
ARIMA	1
Holt-Linear (extrapolation model)	1
Multiple linear regression (causal)	1
Total	**18**

Overall, ANNs scored the lowest error statistics, suggesting that they are usually the most accurate method. However, while ANNs are an exciting and effective method of forecasting, they are imperfect, and a statistician should consider the nature of the time series, as a different model may outperform an ANN. In some cases, ANNs were prone to overfit the data and for noncomplex linear data, SES or causal models were more likely to be endorsed.

Finally, while time series forecasting methods are often extremely accurate, they may fail to account for extreme events and crises. We consider that our forecasts in the period of interest may be drastically altered by the impact of the COVID-19 pandemic, which radically altered medical practice and hospital capacity in early 2020.