Disease Prevention to Reduce New Hampshire Healthcare Claims and Costs: A Data Mining Approach
Disease Prevention to Reduce New Hampshire Healthcare Claims and Costs: A Data Mining Approach

Rakesh Karn, Rom Khattri, Pradeep Podila, Linda Schumacher
Oklahoma State University, Stillwater, OK

Abstract

The healthcare industry in the United States is going through a paradigm shift, moving away from its focus on treating diseases and toward promoting health, wellness, and preventive public health programs, so that both the individuals and the government can maintain a healthy bottom line. The high-level business problem is to reduce the expected medical costs and number of medical services required by the people of New Hampshire by implementing successful disease prevention programs. The objective is to identify which among the six prevention programs will successfully improve the health of the residents of New Hampshire over nine future years (2012–2020). The business scenario of the case is to identify the preventive programs that are most effective in reducing the costs in New Hampshire and to invest the money in those programs so that the overall health-care overhead costs can be reduced or controlled. The effectiveness of implementing the preventive programs was evaluated using SAS® Enterprise Guide® 5.1 and SAS Enterprise Miner™ 12.1. Time series analysis, in particular, forecasting, is used to project the future health-care services and costs for the years from 2012 to 2020.

Objective

The objective of this analysis was to forecast the reduction in the number of medical services and costs associated with the implementation of five possible disease prevention programs in the state of New Hampshire. The disease groups under consideration were diabetes, heart disease, HIV, lung cancer and trauma. The results of implementing the prevention programs individually or concurrently were evaluated.

Methods

The data were obtained from the Analytics 2013 Data Shootout committee. Three spreadsheet files contained NH population information: 2000 US Census figures, 2010 US Census figures and NH birth and death rates. Medical and pharmacy claims information for the years 2006–2010 were provided in SAS data sets. Relationships between the data files were identified by the service id, diagnostic code, age range, county and gender. The data were manipulated to obtain a dataset of services related to the five disease groups. The dataset contained one record with the total medical cost per service. A second record with the total pharmacy costs was present if pharmacy costs were found related to the medical service entry. The number of trauma services was far larger than any other. The median cost per service was highest for lung cancer and heart disease.

The future medical and pharmacy claims were dependent on population changes in New Hampshire. Interannual population estimates were made using arithmetic and geometric methods. Population projections for the years 2011–2020 were made by Hamilton-Perry and cohort survival methods. After comparing the projections with US Census bureau estimates the arithmetic-interpolation method for interannual years and the cohort component method for future predictions were selected.

Time series analysis was used to project the future health care services and costs. The population estimates and claims data were used to calculate the incidence rates by disease group, age range, county and claim type. SAS® Enterprise Miner™ 12.1 time series nodes were used to forecast the future incidence rates, using service per_capita, as the target variable. The candidate exponential smoothing techniques were double, linear, and damped trend smoothing. A second time series node using cost_per_service as the target variable forecast costs using the same smoothing candidate techniques.

Results

The estimates of the future costs by disease group and cohort were calculated using the most likely forecast figures. The improvement impact of the six prevention programs was given as the percent reduction in propensity. These percent reductions and most likely forecast number of services were used to calculate the reductions in the number of future services.

Conclusions

Because of synergy or interaction effects, implementing all programs concurrently provided the greatest cost savings. The health prevention programs under consideration will benefit the people of New Hampshire. It is recommended that all programs are implemented and run concurrently. The minimum anticipated savings in costs is approximately $572,111 or 3.3% of the expected baseline cost of $17,279,931. This amount should be used as the initial funding for the prevention programs being run concurrently.

References

- Improve Predictive Models - And Decision Making - By Combining Time Series Analysis And Data Mining. SAS.com magazine.
- Scott RD, Solomon SL, McGowan JE. Applying Economic Principles to Health Care. CDC and Emory University, Atlanta, GA.
- Tang G. Test and Data Mining to Investigate Expenditures on Prescribed Medicines. University of Louisville, KY.
Data Exploration and Preparation

Identify prevention programs that will improve the health of the residents of New Hampshire.
- BMI reduction
- Stop smoking
- Cholesterol control
- Safety program
- Drug rehab
- Safe sex programs
- Improvement measured by forecasted reduction in cost and services

Integrate
- Merge datasets by key variables
- Select medical claims by ICD-9 codes
- Match medical & pharma claims by service ID
- Pharmacy Refills for chronic conditions
- Filter Negative Cost
- Cutoff Cost at Extreme Percentiles

<table>
<thead>
<tr>
<th>SAS Dataset</th>
<th>Records</th>
<th>Services</th>
<th>Common Service ID</th>
<th>Records Common Service ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical Claims</td>
<td>1,906,287</td>
<td>1,800,186</td>
<td>185,270</td>
<td>185,270</td>
</tr>
<tr>
<td>Pharmaceutical Claims</td>
<td>1,927,153</td>
<td>1,791,002</td>
<td>266,911</td>
<td>266,911</td>
</tr>
</tbody>
</table>

Refills - Pharmaceutical Claims Only
- PROC SQL algorithm to find most frequent disease group per first word of drug name
- Heart Disease Pharma Claims
- Diabetes Pharma Claims
- Caveat – No county identification

Medical and Pharmacy Claims

Number of Services by Disease Group

Box Plot Comparison of Service Cost by Disease Group
Population Estimation

![Population Graphs]

- **2000–2010 Intercensal:** Arithmetic, Geometric
- **2011–2020 Projections:** Cohort Survival, Hamilton Perry
**Time Series Exponential Smoothing**

Forecasting with Time Series – Exponential Smoothing Node

- **Cross sectional variables**: disease group, age, county, claim type: medical or pharmaceutical
- **Model candidates**: double, linear, damped trend
- **Model selection statistic**: mean square error

**Target**: services per capita

**Interval**: Year

**Accumulation**: Total

**Smooth Outliers with predicted value**

**Forecasting Method**: Best

Baseline Services: 2,178,649

Concurrent Services Reduction: 110,510

5.1% reduction in number of services

Baseline Cost: $17,279,931

Concurrent Programs Savings: $572,111

3.3% reduction in cost

Comparison of Time Series Forecasts