Utilising neural network applications to enhance efficiency in the healthcare industry: predicting populations of future chronic illness

Stephan Kudyba*

Department of Management,
New Jersey Institute of Technology,
University Heights, Newark, NJ
E-mail: Stephan.P.Kudyba@NJIT.edu
*Corresponding author

G. Brent Hamar and William M. Gandy

Informatics, Healthways Inc.
E-mail: Brent.Hamar@Healthways.com
E-mail: Bill.Gandy@Healthways.com

Abstract: Advanced analytic and forecasting methodologies can enable organisations to more fully leverage the data resources available to them. In the healthcare industry, service providers can use data mining methods to enhance the decision-making process in optimising resource allocation by identifying the sources of future high-cost treatment in a given health plan population. The following paper includes a case study by Healthways Inc. that illustrates how predictive modelling techniques (e.g., neural networks) can help healthcare providers identify the sources of future high resource demand, enabling them to more effectively apply preemptive treatment to mitigate future high-cost treatment of fully developed cases of chronic illness.

Keywords: predictive modelling; data mining; neural networks; healthcare management; decision support systems.

Reference to this paper should be made as follows: Kudyba, S., Hamar, B. and Gandy, W.M. (xxxx) ‘Utilising neural network applications to enhance efficiency in the healthcare industry: predicting populations of future chronic illness’, Int. J. Business Intelligence and Data Mining, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Stephan Kudyba is a Faculty member in the department of management at New Jersey Institute of Technology where he teaches courses in the graduate program on Data Mining and Knowledge Management and enhancing general corporate productivity through IT and strategic initiatives. He has written/edited three books Data Mining and Business Intelligence: A Guide to Productivity, IT Corporate Productivity and the New Economy and Managing Data Mining: Advice from Experts.

G. Brent Hamar is Director of Informatics at Healthways Inc. and has conducted epidemiological research studies, public health-related projects, health services research and corporate healthcare analytic activities. He has a Masters Degree in Public Health from University of Alabama and a DDS from the University of Tennessee Center for Health Sciences.

Copyright © 200x Inderscience Enterprises Ltd.
1 Introduction

Organisations in the healthcare industry have intensified their initiatives to enhance the efficiency of their operations through effective resource allocation. Given the increased level of competition in today’s digital, information economy, organisations are faced with the task of increasing productivity by more efficiently allocating available resources in producing goods and services to meet the demands of their customers. The healthcare industry has been faced with a number of additional factors that have increased the complexity of managing available resources. Some of these include an increase in the aging population, costs for defensive medicines, optimising existing healthcare facility usage (e.g., staffing doctors and nurses along with designated bed utilisation rates) and the introduction of new organisations such as HMOs and PPOs (Smith-Daniels et al., 1988).

One way to enhance operational efficiency in this sector is by more accurately identifying the sources of resource demand by focusing on the needs of specific customer segments and initiating strategic healthcare management to better allocate available resources in meeting those demands (McLaughlin et al., 1995; Heskett, 1983). More formal analytic techniques such as stochastic trees have been utilised to help increase operational efficiencies by enhancing the decision-making process in medical treatment procedures (Hazen, 1992, 2000). Other analytic methodologies involving data mining techniques, enable decision makers to identify patterns in clinical, claims and activity-based historical data, to better understand explanatory relationships in the data and create models to more accurately predict future resource demand (Xiaohua, 2005). Artificial neural networks are computer algorithms that identify relationships in historical data that can be used for classification and prediction (Bishop, 1995; Swingler, 1996). Reducing the uncertainties in process resource requirements through enhanced predictive capabilities is seen to increase efficiency across industry sectors (Kudyba and Hoptroff, 2001). More specifically in the healthcare sector, increased predictive capability helps reduce inefficient allocation of resources in health services as it decreases the variability in treatment processes which is associated with the complexity of health cases (Yang, 1992). For example, by more accurately identifying patients in the high-risk category of incurring long term, chronic illness, healthcare organisations can incorporate preemptive treatment procedures to help mitigate the future extensive and costly allocation of resources to treat fully developed chronic illnesses.

The purpose of this paper is to illustrate the benefits of incorporating predictive modelling to more accurately identify those patients who are likely to develop chronic illness. This enables service providers to better allocate resources in applying preemptive treatment to deter the development of these illnesses to extreme, fully developed states. The result is a reduction in utilisation of available resources and associated costs in
treating fully developed illnesses. This work incorporates a case study conducted by Healthways Corporation, a leading provider of healthcare products and services.

2 Conceptual issues (Healthways’ Study)

Achieving an optimum resource allocation of people, technology, and general facilities to effectively treat patients who are at a high risk for future medical illness is essential in improving the general health of patients, while also reducing healthcare costs. Efficiently designed programs seek to positively influence patient behaviours, as well as their treatment compliance rates. This, in turn, results in healthier patients who are at lower risk of hospitalisation, emergency room visits or exacerbation of chronic disease conditions (Sidorov et al., 2002).

Typically, a relatively small percentage of a health plan’s members account for a disproportionately large percentage of total healthcare costs (Shelton, 2002; Woo and Cockram, 2000; Ash et al., 2001). Consequently, there is tremendous value to be derived from correctly identifying high-risk patients before their health status begins to significantly degrade, and offering them the appropriate behavioural or clinical interventions.

In order to identify those patients who would benefit the most from disease management and educational efforts, many health plans ‘risk stratify’, or classify their members who suffer from chronic disease conditions. For example, a health plan might stratify their diabetes members into three groups based on estimated current and/or future risk (high, medium, and low risk). In most cases, insurers rely on rules-based risk stratification models to classify members within a given disease population.

However, there are limitations to rules-based stratification algorithms that diminish their accuracy with respect to predicting future risk. The first, and perhaps most prominent, is that rules-based methods are fairly subjective as there exist no widely accepted clinical standards for rating disease specific or total population risk. Second, rules-based algorithms are frequently created from a loose combination of historical experience and principles taken from the medical literature, rather than being systematically derived and validated through inclusive, empirical methods. The principal implication of these limitations is that rules-based stratification methods tend to both lack reasonable predictive power, and exhibit regression to the mean with repeated use (Cousins et al., 2002).

Rules-based approaches have been shown to produce positive results for certain applications however. These types of approaches are very good when there is a clear delineation between choices. For example, certain biometric data (e.g., blood pressure, heart rate and body mass index) received from a patient suffering from Congestive Heart Failure (CHF) might lead a physician to reach the same conclusion, time over time. These types of situations are optimal for rules-based systems. In cases where data are non-linear, they require combining information from multiple data types, and may not contain valid distributions within factors or variables of interest, rules-based systems alone seldom produce the best possible results. Consequently, data mining techniques allow for the development of models that allow for the combination of multiple linear and non-linear techniques, thereby ensuring the production of the most desirable outcomes. Models developed through properly applied data mining techniques have been shown to outperform models derived from rules-based approaches alone. The reason for this is
simple: seldom does the entire distribution associated with a naturally occurring factor or variable have a perfect association with a rule or model. Rules-based approaches have difficulty dealing with variability or exceptions. Models developed through data mining techniques are optimised to take into account variability or exceptions.

3 Neural networks

A prominent methodology in the data mining, data analysis realm is the neural network approach. The neural network methodology is the utilisation of complex computer algorithms that create models that explain variance in historical data. One of the advantages of this method is the ability to identify both linear and non-linear patterns that exist among data variables, which is achieved through the incorporation of the Multi-layer Perceptron technique (Rumelhart and McClelland, 1986). The neural network architecture simply creates weights for input data variables to best explain the variance in dependent variables in historical sample data. This weight adjustment process is conducted through the incorporation of a back-propagation testing algorithm (Hinton, 1992). Neural networks are often discussed in context of attempting to mimic the learning and decision-making process of the human brain (e.g., artificial intelligence), since the architecture is composed of interconnected entities, which are compared to neurons of the brain. Neural networks are widely incorporated in the realm of data analysis to enhance predictive capabilities in a variety of business and scientific applications.

4 Methodology

4.1 Sample and data

Given the unreliable results associated with rules-based risk stratification models, HWAY uses an empirical claims-based approach to predictive modelling and risk stratification. The purpose of the HWAY Initiative is to create a model that correctly identifies future high-risk patients based on information derived from a health plan’s claims history. High risk is defined as those patients who are most likely to exhibit high levels of utilisation and/or medical expenses during a future 12-month outcome period.

HWAY collects and processes all relevant medical claims, pharmacy, lab result and clinical data to develop, calibrate, and implement predictive modelling risk factors for an entire population (usually, 50–75 factors are created from those sources to predict high-risk patients). The independent variable risk factors are based on epidemiological and clinical experience with chronic disease conditions, as well as administrative experience with a wide range of commercial health plans. Typically, a health plan provides two or three years worth of claims data from which these risk factors are extracted. After a full set of risk factors have been extracted from client claims data, HWAY utilises a wide array of artificial intelligence-based, statistical and econometric methods to investigate the data and create a comprehensive predictive model. In general, a study methodology follows these basic steps:
Utilising neural network applications

• define a study population of continuous enrollment for nine consecutive quarters of plan membership
• extract from the data set, the total population risk factors from Year One and outcomes from Year Two
• train and develop a neural network model that predicts Year Two outcomes as a function of Year One factors
• apply the final tested and cross-validated neural network model to predict the Year Three high-cost members
• compare actual Year Three outcomes to Year Three predictions
• revalidate the model against prior period results to ensure robustness.

This study was a retrospective cohort analysis on members of a large commercial health plan located in the southeastern USA. Two years of claims data were used to develop and calibrate the model, using split-sample validation to ensure the reliability (i.e., repeatability) as well as the validity of this predictive model. Cost, rather than utilisation, was used as the dependent measure, because it is often more sensitive to the overall severity associated with medical services. Predictors (i.e., independent variables selected as a sub-set of all risk factors extracted) were selected based on their predicted probabilities using decision trees. Decision Trees are commonly used to select input variables for predictive models when the input-to-target relationship is non-linear, non-additive or unknown. A ‘selection tree’ (i.e., a decision tree used for variable selection) was used to identify variables with high predicted probabilities for inclusion in the neural network model. Variables were retained based on a Chi-Square Test with a significance level of 0.2. Those variables with the highest predicted probabilities were selected for inclusion in the neural network predictive model.

Two years of claims data from a large commercial health plan were used to build this model (May 1997–July 1999). A third year of claims data was used to assess the accuracy of this model (CY 2000). Professional, Pharmacy, and Institutional claims for continuously enrolled members in this health plan’s HMO and Point of Service (POS) products were provided for approximately 55,000 total members. Professional and Pharmacy data were available on the claim level when Institutional claims were reported as a complete episode of care.

4.2 Variables

Individuals were randomly selected for model development. However, for study inclusion, these members had to be continuously enrolled for a minimum of nine consecutive quarters. Previous research has supported the use of nine consecutive quarters of claims data partially owing to administrative problems associated with medical claims processing (Ridinger and Rice, 2000). Using less data can affect the ability to validly represent full information contained within the medical claims, as well as potentially skew the predictive outcomes (as a result of missing or incomplete data). Relevant factors were extracted from the first four quarters (Year One) of available data, which included a total of 59 variables. The dependant variable was extracted from the last four quarters of data (Year Two). Our dependent variable, high cost, was transformed
into a binary outcome based on the top 15% of total medical expenditures. Note that no data were extracted from quarter five – this convention adjusts the data set for any claims lag associated with late reporting or other administrative delay. Claims lag refers to the time period between the medical service date and the medical claims processing/payment date. It is common to find a six-to-eight week claims lag between service and payment dates, obviously affecting a researcher’s ability to accurately identify all completed medical procedures using a health plan’s claims data.

In general, the predictive model was produced through the following steps:

- 59 variables were extracted from claims data
- a variety of methods, including selection trees, logistic regression models and principal component analysis, were used to prune the number of independent variables used by the neural network model
- a neural network processed 14 of the 59 variables to produce the best predictive model (see Figure 1).

**Figure 1** The neural network used in this study included 14 input variables, one hidden layer, three hidden units and one target variable. The model architecture was a multi-layer perceptron, using backward propagation with randomised target bias weights. Note that the figure below does not illustrate direct connections.
Results of the model were evaluated using the following steps:

- Receiver Operating Characteristics (ROCs) and Sensitivity Curves were generated on a Year One to Year Two cross-validation (i.e., withheld) data set
- the model was applied to all individuals in the Year Two data in order to predict membership in the Year Three high-cost class
- all individuals in the Year Two data set received a Predictive Risk Score for Year Three
- all Year Two members were rank ordered from highest to lowest, based on their predictive risk score for a Year Three outcome
- actual Year Three results were compared to the results obtained from the predictive model.

5 Study results

A total of 54,206 health plan members from 3 to 71 years of age met the continuous Year One to Year Two enrollment criteria required for model development and calibration. In Year Three, only 46,141 of the 54,206 continuously enrolled members actually incurred claims. Table 1 contains the descriptive statistics associated with the 14 Risk Factors used in the neural network predictive model. The average age for these members was 34 years, with a gender distribution of 51.2% females and 48.8% males. Roughly 4% of this population was diagnosed with diabetes, and the average total annual cost for these members was $1488.50.

5.1 Year One to Year Two results

Results were evaluated by applying ROCs and Sensitivity Curves (Hanley and McNeil, 1982), which plots the sensitivity of the model vs. the false positive fraction (1–Specificity). The area under the curve is reported as a C-statistic and reflects the model’s overall accuracy. As illustrated in Figure 2, the Year One to Year Two ROC equals 0.807 (0.794–0.819), (asymptotic 95% confidence interval), reflecting very good model performance. In simplest terms, this model was approximately 81% accurate in categorising patients as being high versus low cost using Year One data to predict Year Two cost.

The Sensitivity Curve in Figure 3 illustrates the true positive capture (i.e., sensitivity) at various screening thresholds. The sensitivity analysis of Year Two data shows a high true positive capture rate. A screening threshold of 10% is associated with 42% true capture rate. In other words, an expected true capture rate of 42% of Year Two future high-cost patients would be expected from intervening with only the top 10% of patients predicted by this model as being future high cost. Likewise, screening thresholds of 25%, 50% and 80% are associated with true capture rates of 66%, 85% and 97%, respectively.
Table 1  Descriptive statistics for the input variables used in the predictive model

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>34.0</td>
<td>16.4</td>
<td></td>
</tr>
<tr>
<td>Gender (% Female)</td>
<td></td>
<td></td>
<td>51.2</td>
</tr>
<tr>
<td>Diagnosed with diabetes</td>
<td></td>
<td></td>
<td>3.95</td>
</tr>
<tr>
<td>The change in paid physician dollars over time^2</td>
<td>590.5</td>
<td>1513.9</td>
<td></td>
</tr>
<tr>
<td>The total claims for all drugs</td>
<td>8.6</td>
<td>13.1</td>
<td></td>
</tr>
<tr>
<td>The change in total drug claims over time^3</td>
<td>0.4</td>
<td>3.7</td>
<td></td>
</tr>
<tr>
<td>The number of unique drug claims</td>
<td>4.2</td>
<td>5.0</td>
<td></td>
</tr>
<tr>
<td>Total paid for outpatient dollars</td>
<td>710.2</td>
<td>2156</td>
<td></td>
</tr>
<tr>
<td>Total paid for pharmacy dollars</td>
<td>334.6</td>
<td>705.5</td>
<td></td>
</tr>
<tr>
<td>The change in paid pharmacy dollars over time^2</td>
<td>23.9</td>
<td>205.5</td>
<td></td>
</tr>
<tr>
<td>Somatic diagnostic codes^3</td>
<td>0.12</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>Somatic provider codes^4</td>
<td>0.56</td>
<td>3.6</td>
<td></td>
</tr>
<tr>
<td>Year one total dollars</td>
<td>1488.5</td>
<td>4424.8</td>
<td></td>
</tr>
<tr>
<td>The number of unique ICD9-Codes</td>
<td>12.0</td>
<td>15.9</td>
<td></td>
</tr>
</tbody>
</table>

N = 54,206.

^1 All factors were extracted from the first four quarters (Quarters 1–4) of available data (i.e., Year One).

^2 All changes in dollars and claims over time were calculated as the difference between (Q3 + Q4) and (Q1 + Q2).

^3 HWAY-derived ICD-9 codes based on common somatisation disorders.

^4 HWAY-derived Physician Specialty codes based on common somatisation disorders.

Figure 2  ROC curve for Year One data predicting Year Two outcomes. This model was approximately 81% accurate in identifying future high-cost patients.
5.2 Year Two to Year Three Results

As illustrated in Figure 4, the ROCs Curve for Year Two to Year Three data also shows very good model performance; ROC = 0.827 (0.815–0.838). This model accurately identified approximately 83% of future high-cost patients using Year Two data to predict Year Three cost, which was slightly better than the Year One to Year Two results.

Figure 4 ROC curve for Year Two data predicting Year Three outcomes. This model was approximately 83% accurate in identifying future high-cost patients.

Similarly, Figure 5 illustrates the Sensitivity Curve associated with Year Three data, which demonstrates a high true positive capture rate. Screening thresholds of 10%, 25%, 50% and 80% are associated with true capture rates of 41%, 66%, 86% and 97%, respectively.
Figure 5  The Sensitivity Curve illustrates the true positive capture (i.e., sensitivity) at various screening thresholds. The Sensitivity analysis of Year Two to Year Three data demonstrates a high true positive capture rate.

5.3 Year Two vs. Year Three results

Figure 6 illustrates both Year Two and Year Three Sensitivity Curves, and Table 2 contains meaningful threshold points along these curves. The neural network predictive model produced very similar results, in terms of number of true positive captures, for Year Two as well as Year Three data. That is, the model trained on Year One to Year Two data demonstrated comparable results for Year Two and Year Three. This result demonstrates the overall high external validity and repeatability (i.e., reliability) of the predictive model.

Figure 6  Year Two and Year Three Sensitivity Curves. As illustrated, the neural network predictive model produced very similar results for both years demonstrating the accuracy and robustness of this model.
Utilising neural network applications

Table 2  Meaningful threshold points for Year Two and Year Three sensitivity analyses

<table>
<thead>
<tr>
<th>Screening threshold (%)</th>
<th>Year Two true positive captures (%)</th>
<th>Year Three true positive captures (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td>17</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>10</td>
<td>42</td>
<td>41</td>
</tr>
<tr>
<td>15</td>
<td>51</td>
<td>51</td>
</tr>
<tr>
<td>20</td>
<td>59</td>
<td>59</td>
</tr>
<tr>
<td>25</td>
<td>66</td>
<td>66</td>
</tr>
<tr>
<td>30</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>40</td>
<td>78</td>
<td>79</td>
</tr>
<tr>
<td>50</td>
<td>85</td>
<td>86</td>
</tr>
<tr>
<td>60</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>70</td>
<td>94</td>
<td>95</td>
</tr>
<tr>
<td>80</td>
<td>97</td>
<td>97</td>
</tr>
</tbody>
</table>

6 Conclusions

This study demonstrates that neural network predictive models accurately identify a large percentage of patients who are at high risk for future medical cost. In general, the ROC curves support the assertion that there is approximately 79–84% probability of correctly identifying true future high-risk patients (95% Confidence Level) from these data. In Signal Detection Theory, the area under the ROC curve corresponds to the probability of correctly identifying 'noise' and 'signal plus noise'. Sensitivity analyses revealed that the neural network predictive model demonstrated a high-level of predictive validity. For example, by intervening with only the top 10% of patients predicted by the model as being future high risk, we would expect to capture at least 42% of the future true high-risk patients. Likewise, intervening with 50% of this population would capture at least 85% of the future true high-risk patients. Furthermore, this study also demonstrates the accuracy and robustness of using neural network predictive modelling techniques on claims data to predict future high-risk patients as results from both ROC and Sensitivity Curve analyses demonstrate a high level of predictive validity.

It is possible to develop neural network models, which identify considerably more future high-risk patients than many traditional rules-based or regression-based models (Kiernan et al., 2001). The positive impact in the efficiency and cost effectiveness of care management intervention programs on at-risk health plan members is significant. Prospectively, identifying high-risk patients and subsequently adjusting healthcare management interventions based on patients’ risk stratification levels can result in improvements in the quality of care received by these patients, as well as positively impacting associated healthcare costs (Lieu et al., 1998). Predictive models such as this can help to focus resources on the highest risk patients in the population, resulting in potentially more effective healthcare treatment for those patients, with the intention of mitigating the number of fully developed chronic illness cases in the future (Kudyba et al., 2005).
The predictive model results reported in this paper are based on a specific objective for a defined population at a given point in time. HWAY recognises the value of predictive modelling and continues ongoing research with predictive models for positive healthcare outcomes through early identification and intervention.

Acknowledgement

Special thanks to Dr. Thad Perry for his help regarding this project.

References


Utilising neural network applications


