October 2005

Comparative effectiveness of total population versus disease-specific neural network models in predicting medical costs

Albert G. Crawford
Thomas Jefferson University

Joseph P. Fuhr Jr.
Thomas Jefferson University

Janice Clarke
Thomas Jefferson University

Brandon Hubbs
American Healthways, Inc.

DISEASE MANAGEMENT
Volume 8, Number 5, 2005

Comparative Effectiveness of Total Population versus Disease-Specific Neural Network Models in Predicting Medical Costs

ALBERT G. CRAWFORD, Ph.D., M.B.A., M.S.I.S.,[1] JOSEPH P. FUHR, JR., Ph.D.,[1]
JANICE CLARKE, R.N., B.B.A.,[1] and BRANDON HUBBS, M.A.[2]

ABSTRACT

The objective of this research was to compare the accuracy of two types of neural networks in identifying individuals at risk for high medical costs for three chronic conditions. Two neural network models—a population model and three disease-specific models—were compared regarding effectiveness predicting high costs. Subjects included 33,908 health plan members with diabetes, 19,264 with asthma, and 2,605 with cardiac conditions. For model development/testing, only members with 24 months of continuous enrollment were included. Models were developed to predict probability of high costs in 2000 (top 15% of distribution) based on 1999 claims factors. After validation, models were applied to 2000 claims factors to predict probability of high 2001 costs. Each member received two scores—population model score applied to cohort and disease model score. Receiver Operating Characteristic (ROC) curves compared sensitivity, specificity, and total performance of population model and three disease models. Diabetes-specific model accuracy, C = 0.786 (95%CI = 0.779–0.794), was greater than that of population model applied to diabetic cohort, C = 0.767 (0.759–0.775). Asthma-specific model accuracy, C = 0.835 (0.825–0.844), was no different from that of population model applied to asthma cohort, C = 0.844 (0.835–0.853). Cardiac-specific model accuracy, C = 0.651 (0.620–0.683), was lower than that of population model applied to cardiac cohort, C = 0.726 (0.697–0.756). The population model predictive power, compared to the disease model predictive power, varied by disease; in general, the larger the cohort, the greater the advantage in predictive power of the disease model compared to the population model. Given these findings, disease management program staff should test multiple approaches before implementing predictive models. (Disease Management 2005;8:277–287)

INTRODUCTION

ROUGHLY one half of all health care costs in the United States stem from chronic diseases, and this proportion is expected to increase as the proportion of seniors rises.[1] By the year 2020, it is expected that approximately 50% of Americans (157 million) will have one or more chronic illnesses.[2] Each of the three chronic conditions included in this study generates more than $10 billion annually in U.S. healthcare costs:

[1]Department of Health Policy, Jefferson Medical College, Philadelphia, Pennsylvania.
[2]Informatics, American Healthways, Inc., Nashville, Tennessee.

- Diabetes generated $132 billion in direct and indirect costs in 2002.
- Asthma generated $12.7 billion in healthcare costs in 1998.
- Heart failure generated $22 billion in direct costs in 2003.[3]

Over the past decade, disease management (DM) programs have proven effective in controlling these costs while improving health outcomes.[4] DM is a process by which health plan members at risk of chronic illness are identified and targeted for interventions aimed at improving their clinical outcomes, thereby reducing medical costs associated with poorly controlled conditions.[5] A core component of DM is identification of individuals at risk of development or exacerbation of illness and concomitant costs.[5–10] Using various disease markers, types of health services utilization, and healthcare cost levels, predictive modeling techniques have been developed to pinpoint individuals at risk for adverse health outcomes. Predictive models can be either generic or specific in their population of interest. Another dimension of differentiation among predictive models is the specific analytical technique employed, eg, linear or logistic regression analyses, classification/decision trees, or neural networks.[11–13]

Neural network techniques are derived from theories of human cognition and employ nonstatistical algorithms to explain or predict variations in data. In neural networks, the individual inputs to a neuron, with initial values of 0 or 1, are multiplied by their respective weights, and these weighted inputs are summed and processed through a threshold function to determine whether the summed input exceeds the threshold for the neuron.

There are various types of neural networks, based on number of neurons, number of layers, number of hidden layers, and number of outputs. The most common implementation of neural networks is backpropagation. Backpropagation is a two-stage process, consisting of (1) feed-forward activation from the input layer to the output layer, and (2) propagation of errors and adjustments backward to the input layer. In backpropagation, if an output is correct, no change is necessary; if there is a false positive or false negative error, each weight is adjusted according to the direction and degree of error. Backpropagation requires hidden layers, ie, middle layers that provide an internal model of how inputs are related to outputs. As the number of hidden layers increases, the training error rate decreases as a result of increased flexibility in modeling the data.

Applications of neural networks include games, speech synthesis and interpretation, and signal processing, cleaning, and interpretation, as well as the focus of this analysis, medical decision-making, specifically diagnosis. Neural networks have recently been used to predict, among other outcomes, acute pancreatitis patient outcomes,[14] length of stay in a postanesthesia care unit,[15] breast cancer survival,[16] and 5-year colon carcinoma survival.[13]

One important issue is how neural network techniques compare in predictive power with statistical techniques, ie, linear and logistic regression. The four studies cited immediately above all compared neural network analyses with other analytic techniques. In the study predicting acute pancreatitis patient outcomes by Keogan et al,[14] neural network predictions were not significantly better than linear discriminant analysis predictions ($C = 0.83$ and $C = 0.82$, respectively). On the other hand, in the study of postanesthesia care unit length of stay by Kim et al,[15] a neural network predicted correctly in 81.4% of situations, while logistic regression analysis predicted correctly in 65.0%. Similarly, the study of breast cancer survival by Burke et al[16] found that both a backpropagation neural network ($C = 0.768$) and a probabilistic neural network ($C = 0.759$) were significantly more accurate than the pTNM (primary tumor, regional lymph nodes, and distant metastases) staging system in predicting breast cancer survival ($C = 0.720$). Finally, the comparison by Snow et al[13] predicting 5-year colon carcinoma survival found that a neural network performed better than a standard parametric logistic regression in terms of both C-statistics and specificities at 95% sensitivity.

Among the reasons why neural networks often perform better than statistical techniques are the assumptions required by the latter. Clinical research must address numerous di-

chotomous outcomes, including health versus illness, receiving a service or not, being admitted to a hospital or not, and survival itself. While linear regression cannot handle dichotomous outcomes, logistic regression can do so. Still, no regression technique has the capability to handle associations between outcomes, continuous or dichotomous, and predictor variables whose effects are neither simply linear nor linear following adjustment through a mathematical transformation, eg, logarithmic/exponential, polynomial, trigonometric. In contrast, neural networks have no such limitations in terms of either forms of variables or forms of associations between variables. While neural networks have the aforementioned strengths in incorporating complex variables and associations and in overall predictive power, they also have weaknesses. Among their disadvantages are use of hidden layers (which are inherently somewhat indescribable), lack of easily interpretable results, and, specifically, lack of detailed or summary quantitative results.[17,18]

The general aim of this analysis was to compare the effectiveness of two neural network modeling approaches in predicting high medical costs: a model based on a population of health plan members versus three cohort models targeting members with specific diseases, ie, asthma, diabetes, and cardiac conditions—congestive heart failure (CHF) and coronary artery disease (CAD). A more specific aim of the analysis was to determine the more effective prediction method to increase the clinical benefits and cost effectiveness of the DM program.

The conditions studied—diabetes, asthma, and CHF and CAD combined—were selected because they are the three conditions most commonly targeted by DM programs. Of all health plans participating in the 2001 American Association of Health Plans Annual Industry Survey, 97% had a diabetes DM program, 86% had an asthma DM program, and 83% had a CHF DM program.[19]

The total population model was developed through analysis of a health plan population containing members with and without chronic diseases. Each disease-specific model was developed by focusing on the cohort of members diagnosed with or having risk factors for that disease. The choice of the outcome of high future costs was driven by two exigencies: (1) the availability and importance of high costs as an indicator of adverse health status, and (2) the goal of managing the DM program as cost effectively as possible.

MATERIALS AND METHODS

The total study population consisted of 375,426 members of a health plan where American Healthways, Inc. provided DM services. The three disease cohorts included 33,908 members with diabetes, 19,264 members with asthma, and 2,605 members with CHF and/or CAD. A caveat is in order regarding the disease hierarchies in the three cohorts. On the one hand, the asthma cohort is relatively uniform: only 0.8% of its members have cardiac conditions and only 0.3% have diabetes; moreover, since there is a separate chronic obstructive pulmonary disease (COPD) cohort not included in these analyses, there are no members with COPD in the asthma cohort. On the other hand, the diabetes and cardiac cohorts are more complex: in the diabetes cohort, 18.2% have cardiac conditions and 9.6% have asthma; and, in the cardiac cohort, 5.6% have diabetes and 11.3% have asthma.

The time frame of the study was calendar years 1999–2001, where 1999 was modeling year 1, 2000 was modeling year 2, and 2001 was the evaluation year.

The outcome variable predicted by the models was high medical costs, defined as the top 15% of the total cost distribution, ie, the segment targeted for the most intense disease management interventions. This kind of operational definition of high costs reflects the fact that approximately 10% of the US population is responsible for roughly 70% of direct medical costs.[20] A set of potential predictive risk factors was identified for use in modeling both the total population and each of the three disease cohorts. American Healthways, Inc. collected and processed all relevant medical (inpatient, outpatient, and physician) and pharmacy claims, laboratory results, and other clinical data to identify risk factors and to develop, calibrate, and implement the predictive mod-

els for this member population. In model development, a set of more than 100 risk factors was compiled based on (1) epidemiological and clinical knowledge of chronic disease conditions and their progression, and (2) the clinical and administrative experience of a wide range of commercial health plans.

Proprietary algorithms employed a variety of factors derived from members' claims histories to identify members with each chronic condition. A specific identification algorithm was developed for each disease, including data drawn from both medical claims—International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9) diagnosis codes, ICD-9 procedure codes, and Current Procedure Terminology (CPT) procedure codes—and pharmacy claims—National Drug Code (NDC) codes. Selection criteria for diabetes and for asthma included a combination of specific ICD-9 and NDC codes; selection criteria for cardiac conditions included a combination of specific ICD-9, CPT, and NDC codes. Given that the identification of each disease cohort required at least two services with specified ICD-9 diagnosis codes, the false positive rate for each disease cohort is less than 5%, suggesting that false positives do not represent a major limitation of these analyses.

Decision trees were used to reduce the total set of more than 100 predictive factors to smaller subsets of factors based on their relationships to the outcome variable, high medical costs. The threshold to include a factor in each subset was a chi-square significance test where $p < 0.20$. Each of the four neural network models contained one hidden layer, and within this layer a default value of three hidden units was used. The average error model selection criterion was used for each neural network. All classification and neural network modeling was performed using SAS Enterprise Miner,® version 4.1.[21] The ultimate criterion for inclusion in the model was that the importance statistic, representing the relative importance of a variable and generated by the decision tree component of SAS Enterprise Miner, had a value greater than or equal to 0.05.[21]

Four metrics by which a predictive model can be evaluated are (1) the true positive rate (sensitivity), (2) the true negative rate (specificity), (3) the false positive rate, and (4) the false negative rate. These rates are interrelated. The false positive rate represents the percentage of members who were predicted to incur high costs but who did not actually incur high costs. The false negative rate represents the percentage of members who were not predicted to incur high costs but who actually incurred high costs.

Receiver Operating Characteristic (ROC) curves were plotted to compare the performance of the total population model with that of each disease-specific model. These ROC curves plot the sensitivity of each model as a function of its false positive rate (1 minus specificity); the area under the ROC curve represents the overall accuracy or performance of each model.[22,23] The three comparisons focused on differences between the confidence intervals around the pairs of C-statistics generated by applying the two predictive models (total population and disease specific) to each disease cohort. The C-statistic can range from 0 to 1; a C-statistic of 0.50 would indicate that a model was 50% accurate in categorizing members as having high versus low costs (odds which are no greater than chance); a C-statistic of 1.00 would indicate a perfect model, ie, a model that perfectly predicts whether members incur high cost or not.

The four neural network models were constructed using claims data for modeling year 1 (1999) to predict which members had high costs in modeling year 2 (2000). After the models were developed and validated, they were applied to modeling year 2 (2000) claims factors to predict year 3 (2001) costs. The outcome of interest is the probability that members have high costs in year 3. Members in each of the three disease categories received two scores, one based on the total population model applied to the disease cohort and the other based on the respective disease-specific model. True positive high cost members were identified as those in the top 15% of the year 3 cost distribution.

The choice of a dichotomous outcome rather than a continuous one was guided by comparative analysis of models with the two forms of outcomes. One population model was fitted to

a dichotomous target, and another population model was fitted to a continuous, ie, interval scale target. Two sets of scores were assigned to the population, and model performance was compared at the 15% screening level. The results were consistent with those derived from analyses of data from other health plans: dichotomous targets consistently perform better than continuous targets, with a significantly higher percentage of high cost members captured at the 15% screening threshold, 56.25% versus 54.19%.

Sensitivity analyses were performed using different predictive modeling scenarios, varying the target percentages defining members' high medical costs, ie, 1%, 5%, 10%, 15%, 20%, and 30%. These analyses allowed for an assessment of any variations in the accuracy of the models depending on the threshold selected.

RESULTS

Total population model

The total population model (Table 1) includes the following predictive factors, listed in descending order according to the magnitude of their importance statistics: total medical costs, physician costs, prescription drug costs, number of unique diagnoses, age, number of prescription drug claims, number of unique procedures, hypertension symptoms, CAD symptoms, inpatient costs, and diabetes symptoms. Four of the 11 factors in the total population model indicate costs, another three are symptoms, three indicate intensity/complexity of utilization, and the remaining factor is age.

Disease-specific models

Diabetes. The diabetes model includes the following predictive factors, shown in descending

Table 1. Importance Statistics for Predictive Factors in Total Population Model and Three Disease-Specific Models

Factor	*Population*	*Diabetes*	*Asthma*	*Cardiac*
Diabetes symptoms	0.0652	0.3885		
Hypertension symptoms	0.2786			0.4095
Coronary artery disease symptoms	0.1283			0.4149
Asthma symptoms			0.4201	0.6126
COPD symptoms				0.6809
Cancer symptoms				0.1592
Osteoarthritis symptoms				0.1104
End stage renal disease symptoms		0.0743		
IBD symptoms				0.2113
Age	0.5066	0.2965	0.6055	0.7193
Total medical costs	1.0000	1.0000	1.0000	0.9387
Inpatient costs	0.1029			0.2112
Outpatient costs		0.4799	0.4774	0.5133
Physician costs	0.9509	0.9134	0.8925	1.0000
Prescription drug costs	0.5305	0.4509	0.4275	
Number of hospitalizations		0.4277	0.1719	
Number of specialist visits			0.3477	
Number of emergency room visits				0.4930
Number of prescription drug claims	0.4949	0.3526	0.4014	
Number of unique diagnosis codes	0.5119	0.2775	0.6834	
Number of procedure codes		0.3537	0.3971	0.6710
Number of unique procedure codes	0.4754	0.2936		

COPD, chronic obstructive pulmonary disease; IBD, inflammatory bowel disease.

order according to their importance statistics: total medical costs, physician costs, outpatient costs, prescription drug costs, number of hospitalizations, diabetes symptoms, number of procedures, number of prescription drug claims, age, number of unique procedures, number of unique diagnoses, and end stage renal disease symptoms. In the diabetes model, four of the 12 factors indicate costs, only two are symptoms (one being diabetes itself), five indicate intensity/complexity of utilization, and the remaining factor is age.

Asthma. The asthma model includes the following predictive factors, listed in descending order of importance: total medical costs, physician costs, number of unique diagnoses, age, outpatient costs, prescription drug costs, asthma symptoms, number of prescription drug claims, number of procedures, number of specialist visits, and number of hospitalizations. In the asthma model, four of the 11 factors indicate costs, only one (asthma) is a symptom factor, five indicate intensity/complexity of utilization, and the remaining factor is age.

Cardiac. The cardiac model includes the following predictive factors, listed in descending order of importance: physician costs, total medical costs, age, chronic obstructive pulmonary disease symptoms, number of procedures, asthma symptoms, outpatient costs, number of emergency room visits, CAD symptoms, hypertension symptoms, inflammatory bowel disease symptoms, inpatient costs, cancer symptoms, and osteoarthritis symptoms. In the cardiac model, four of the 14 factors indicate costs, seven are symptoms, two indicate intensity/complexity of utilization, and the remaining factor is age.

Comparison of total population model with disease-specific models. To compare the effectiveness of the total population model with that of the three disease-specific models, two ROC curves were plotted for each disease. These ROC curves plot the sensitivity of each model versus its false positive rate (1 minus specificity). The area under each curve represents the overall accuracy of that model and is evaluated by the C-statistic. The 95% confidence interval of the C-statistic for each application of the population model is compared with the confidence interval of the C-statistic for the respective disease model. The results are shown in Table 2.

The accuracy of the diabetes model at the 95% confidence level, C = 0.786 (CI = 0.779–0.794) was significantly higher than that of the total population model applied to the diabetic cohort, C = 0.767 (CI = 0.759–0.775). The accuracy of the asthma model, C = 0.835 (CI = 0.825–0.844), was not significantly different from that of the total population model applied to the asthmatic cohort, C = 0.844 (CI = 0.835–0.853). In contrast, the accuracy of the cardiac model, C = 0.651 (CI = 0.620–0.683), was significantly lower than that of the total population model applied to the cardiac cohort, C = 0.726 (CI = 0.697–0.756).

The analyses reported above used the 15% screening threshold, ie, prospectively identifying the top 15% of members in terms of medical costs in the following year. A more fine-grained analysis of the performance of the models at this screening threshold shows that: for persons with diabetes, the total population model (sensitivity = 0.422, specificity = 0.898) performed worse than the diabetes model (sensitivity = 0.475, specificity = 0.907), paralleling

Table 2. Model Performance Comparison: Population Model versus Disease Models

Disease	*Population model accuracy, C, 95% CI*	*Disease-specific model accuracy, C, 95% CI*
Diabetes (n = 33,908)	C = 0.767 CI = 0.579–0.775	C = 0.786 CI = 0.779–0.794
Asthma (n = 19,264)	C = 0.844 CI = 0.835–0.853	C = 0.835 CI = 0.825–0.844
Cardiac (n = 2,605)	C = 0.726 CI = 0.697–0.756	C = 0.651 CI = 0.620–0.683

the results reported above based on C-statistics; for persons with asthma, the total population model (sensitivity = 0.535, specificity = 0.918) and the asthma model (sensitivity = 0.530, specificity = 0.917) were virtually identical, paralleling the C-statistic analyses; and, for cardiac patients the total population model (sensitivity = 0.361, specificity = 0.888) performed better than the cardiac model (sensitivity = 0.302, specificity = 0.877), once again in accordance with the C-statistic analyses.

To summarize, comparing model performance based on the C-statistic demonstrated that the diabetes model performed better than the total population model, the asthma model approximated the total population model in performance, and the cardiac model performed worse than the total population model.

Table 3 shows the results of sensitivity analyses where the threshold percentage at which members were defined as having high medical costs was varied from 1% up to 30%. These comprehensive sensitivity analyses demonstrate the consistency of the patterns reported above, regardless of the threshold percentage selected. The diabetes-specific model is superior to the population model throughout the entire range of screening thresholds. Its advantage in both sensitivity and specificity increases between the 1% and 10% thresholds, plateaus between 10% and 20%, and diminishes somewhat between 20% and 30%, but its superiority is still substantial at the 30% threshold; the diabetes model advantage in sensitivity always exceeds 0.03, and its advantage in specificity always exceeds 0.006. The close similarity between the population model and the asthma-specific model persists across the range of screening thresholds; while there is some fluctuation in the difference in sensitivity, that

Table 3. Sensitivity/Specificity Comparison: Population Model versus Disease Models, by Screening Threshold

	Total population		*Diabetes specific*	
Screening threshold	*Sensitivity*	*Specificity*	*Sensitivity*	*Specificity*
0.01	0.059	0.999	0.059	0.999
0.05	0.203	0.977	0.234	0.983
0.10	0.323	0.939	0.374	0.948
0.15	0.422	0.898	0.475	0.907
0.20	0.502	0.853	0.552	0.862
0.30	0.636	0.759	0.669	0.765

	Total population		*Asthma specific*	
Screening threshold	*Sensitivity*	*Specificity*	*Sensitivity*	*Specificity*
0.01	0.064	1.000	0.060	0.999
0.05	0.256	0.986	0.260	0.987
0.10	0.416	0.956	0.420	0.956
0.15	0.535	0.918	0.530	0.917
0.20	0.623	0.875	0.611	0.873
0.30	0.744	0.778	0.736	0.777

	Total population		*Cardiac specific*	
Screening threshold	*Sensitivity*	*Specificity*	*Sensitivity*	*Specificity*
0.01	0.038	0.995	0.046	0.996
0.05	0.169	0.971	0.153	0.968
0.10	0.269	0.930	0.233	0.924
0.15	0.361	0.888	0.302	0.877
0.20	0.427	0.840	0.358	0.828
0.30	0.568	0.748	0.463	0.729

difference never exceeds 0.012; and the difference in specificity never exceeds 0.002. On the other hand, the disadvantage of the cardiac model compared to the population model, particularly in sensitivity, becomes progressively greater as the threshold increases: the cardiac model disadvantage in sensitivity increases to a maximum of 0.105 at the 30% threshold, and its disadvantage in specificity increases to a maximum of 0.019 at the 30% threshold. Given all of these sensitivity analysis findings, it is reasonable to conclude that the findings based on the 15% screening threshold are not an artifact of the threshold choice but are generalizable to a wide range of plausible thresholds.

DISCUSSION

The primary aim of this study was to compare the relative accuracy, or effectiveness, of a total population neural network model with each of three disease-specific neural network models in predicting which health plan members will incur high costs. The working hypothesis was that the disease-specific models would outperform the total population model. The most striking finding is that of the variable effectiveness of the disease models. When the disease models were compared with the total population model, the diabetes model was more effective than the total population model applied to persons with diabetes, the asthma model was roughly equivalent, and the cardiac model was less effective than the total population model applied to the cardiac cohort. If these results are substantiated by further analyses, they may imply that the choice of approach should be treated as an empirical question to be answered specifically for each disease.

Evaluating these results in light of the DM, predictive modeling, and medical informatics literature is difficult, given little comparable evidence. C-statistics reported in the literature have generally ranged from 0.5 to 0.9. However, valid comparisons of C-statistics require greater similarities in designs and data than can be found in the few published studies. Further research employing more standardized designs is needed to address these issues.

Still, it is noteworthy that the sets of significant predictive factors identified in each of the four models are very similar to those identified in the literature. Typically, prior costs are highly effective predictors of future costs; cost factors taken collectively—including total medical costs, factors which decompose total costs into components (eg, inpatient costs), and changes over time in these factors—represent the most common class of predictors. Intensity/complexity of utilization factors (eg, number of hospitalizations, specialist visits, emergency room visits) and changes therein represent the second most common class of predictors. Symptom factors taken together represent a smaller but important source of predictors, ranging from five symptoms in the total population model to a diagnosis of asthma alone in the asthma model. Finally, age stands alone among the demographic factors as an effective predictor. Its generality of effectiveness is impressive: it has the third highest importance statistic in the asthma and cardiac models, the fifth highest in the total population model, and the ninth highest in the diabetes model.

Looking at the effectiveness of the predictive factors in a more granular way, total medical costs in the prior year was generally the most powerful predictor of high costs; total medical costs had the highest importance in all but the cardiac model, where it had the second highest importance. Total physician cost was the second most powerful predictor in the four models; it had the highest importance in the cardiac model and the second highest importance in the other three models. The third most powerful predictor was age; it was the only powerful demographic predictor, as noted above, and had high importance in all four models.

Five other factors were important predictors in three out of the four models: outpatient costs, prescription drug costs, number of prescription drug claims, number of unique diagnosis codes, and number of procedure codes.

The matrix of effects of symptoms on costs is relatively sparse in that no symptom factor has a high importance rating in more than two of the models, and many of the symptom factors with the highest importance statistics

are somewhat tautological. Hypertension and coronary artery disease are highly important factors predicting high medical costs in the cardiac model. Similarly, diabetes symptoms are highly important in predicting high costs in the diabetic cohort. The fact that asthma symptoms is the only symptom factor predicting high medical costs among persons with asthma highlights the absence of other symptom factors. At the same time, all of these symptom factors except asthma—hypertension, CAD, and diabetes—have high importance ratings in the total population model.

There are some parallels between the results in this study and the results in the literature. A study by Dove et al.[8] sought to identify high-risk members in a subpopulation of managed care organization members with previously low medical costs, employing techniques similar to those reported herein (ie, multiple regression analysis, a comparative study design, and an ROC curve). Dove et al.[7] reported the area under the curve to be 0.73, and found predictors similar to those in the analyses presented herein: diabetes, cardiac, respiratory, and psychiatric conditions (based on medications and diagnoses), "nonhospital, non-emergency department, nonphysician medical claim variable," "composite prescription claim variable: measure of prescription drug classes," and symptoms/comorbidities (truncated at four).

Similarly, a study by Lieu et al[24] aimed at identifying children at high risk of developing asthma rather than identifying individuals with asthma or high medical costs in a broader age range. Their study employed proportional-hazard modeling and their findings regarding predictors parallel the findings reported herein in many respects: "having filled an oral steroid prescription . . . having been hospitalized during the prior 6 months, and not having a personal physician . . . were associated with increased risk of future hospitalization. . . . Classification trees identified previous hospitalization and ED visits, 6 or more β-agonist inhalers (units) during the prior 6 months, and three or more physicians prescribing asthma medication during the prior 6 months as predictors."

The analyses reported herein entail a number of limitations:

- They focus solely on medical cost outcomes. Other outcomes that might be examined include clinical outcomes and indicators of the intensity/complexity of utilization (which were actually employed herein as predictive factors).
- The analyses included data from only one DM program and on only three disease conditions.
- A single modeling technique was employed—neural networks.

An intriguing study finding is the relatively low accuracy of the cardiac model (C = 0.651) relative to the total population model applied to the cardiac cohort (C = 0.726) and relative to the other disease-specific models (asthma model C = 0.835 and diabetes model C = 0.786). One explanation may lie in the relatively small size of the cardiac cohort (n = 2,605) in comparison with the diabetes (n = 33,908) and asthma (n = 19,264) cohorts. Increasing enrollments in the cardiac cohorts of DM programs will permit future testing of this interpretation. Another possible explanation lies in the fact that the cardiac cohort combines members with two conditions, CHF and CAD, increasing the complexity of the cohort and, presumably, reducing the potential for a single model to account simultaneously for the costs of members with either condition.

There is clearly a need for further research to address the questions of the comparative effectiveness of generic versus disease-specific neural network models, and the variable effectiveness of models of different diseases. Incorporating a broader array of disease conditions—other chronic conditions such as cancer, COPD, end stage renal disease, HIV/AIDS, low back pain, and depression, as well as some acute conditions such as high-risk pregnancy—will also be of value.

Testing hypotheses about the variable effectiveness of disease-specific models should include incorporating information about disease progression, and measuring specific disease markers longitudinally over a longer time span, preferably longer than 3 years. In general,

as DM program enrollments and enrollment longevity increase, such data should become more widely available.

CONCLUSION

This study compared the effectiveness of a total population neural network model predicting high medical costs with three disease specific models. The most striking finding of this research is that the effectiveness of predictive models varies by disease, ie, the diabetes model appears to be more effective than the total population model, the asthma model appears roughly as effective, and the cardiac model appears less effective than the total population model applied to the cardiac cohort. If substantiated by further analyses, these results suggest that DM program developers and administrators should test multiple approaches—both generic and disease-specific—before finalizing and implementing predictive models in DM programs. Additionally, the predictive power of a model seems to be directly related to its sample size. Thus, as in applications of statistical analyses in general, DM program developers and administrators should be cautious in applying predictive models to small samples.

ACKNOWLEDGMENTS

This research was supported by an educational grant to Jefferson Medical College from American Healthways, Inc.

REFERENCES

1. Woo J, Cockram C. Cost estimates for chronic diseases. Disease Management and Health Outcomes 2000;8:29–41.
2. Chronic disease prevalence statistics. Partnership for Solutions, Johns Hopkins University, and the Robert Wood Johnson Foundation. Available: ⟨www.chronicnet.org/statistics/prevalence.htm⟩, 2004.
3. Goldfarb N, Weston C, Hartmann C, et al. Impact of appropriate pharmaceutical therapy for chronic conditions on direct medical costs and workplace productivity: a review of the literature. Dis Manag 2004;7:61–75.
4. Shelton P. Disease management programs: the second generation. Disease Management and Health Outcomes 2002;10:461–467.
5. Ash A, Zhao Y, Ellis R. Finding future high-cost cases: comparing prior cost versus diagnosis-based methods. Health Serv Res 2001;36:194–206.
6. Cousins M, Shickle L, Bander J. An introduction to predictive modeling for disease management risk stratification. Dis Manag 2002;5:157–167.
7. Dove H, Duncan I, Robb A. A prediction model for targeting low-cost, high-risk members of managed care organizations. Am J Manag Care 2003;9: 381–389.
8. Grana J, Preston S, McDermott P, et al. The use of administrative data to risk-stratify asthmatic patients. Am J Med Qual 1997;12:113–119.
9. Ridinger M, Rice J. Predictive modeling points way to future risk status. Health Manag Technol 2000; 21:10–12.
10. Sidorov J, Shull R, Tomcavage J, et al. Does diabetes disease management save money and improve outcomes? A report of simultaneous short-term savings and quality improvement associated with a health maintenance organization–sponsored disease management program among patients fulfilling health employer data and information set criteria. Diabetes Care 2003;25:684–689.
11. Trout J, Bishop M. 50 years of successful predictive modeling should be enough: lesson in philosophy of science. Phil Sci 2002;69:S197–S208.
12. Lacson RC, Ohno-Machado L. Major complications after angioplasty in patients with chronic renal failure: a comparison of predictive models. Proc AMIA Symp 2000;91:457–461.
13. Snow P, Kerr D, Brandt J, et al. Neural network and regression predictions of 5-year survival after colon carcinoma treatment. Cancer 2001;91:1673–1678.
14. Keogan M, Lo J, Freed K, et al. Outcome analysis of patients with acute pancreatitis by using an artificial neural network. Acad Radiol 2005;9:410–419.
15. Kim W, Kil H, Kang J, et al. Prediction on lengths of stay in the postanesthesia care unit following general anesthesia: preliminary study of the neural network and logistic regression modelling. Korean Med Sci 2000;15:25–30.
16. Burke H, Rosen D, Goodman P. Comparing artificial neural networks to other statistical methods for medical outcome prediction. IEEE World Congress Comput Intell 1994;4:2213–2216.
17. Adriaans P, Zantinge D. Data mining. Harlow, England: Addison-Wesley, 1996.
18. Weiss S, Indurkhya N. Predictive data mining: a practical guide. San Francisco: Morgan Kaufmann Publishers, Inc., 1998.
19. Welch WP, Bergsten C, Cutler C, et al. Disease management practices of health plans. Am J Manag Care 2004;8:353–361.
20. Berk M, Monheit A. The concentration of health care expenditures, revisited. Health Aff (Millwood) 2001; 20:9–18.

21. SAS Institute Inc. Data mining using Enterprise Miner software: a case study approach. Cary, NC: SAS Institute, Inc., 2000.
22. Hanley J, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Diagn Radiol 1982;143:29–36.
23. Kiernan M, Kraemer H, Winkleby M, et al. Do logistic regression and signal detection identify different subgroups at risk? Implications for the design of tailored interventions. Psychol Methods 2001;6:35–48.
24. Lieu T, Quesenberry C, Sorel M, et al. Computer-based models to identify high-risk children with asthma. Am J Respir Crit Care Med 1998;157:1173–1180.

Address reprint requests to:
Albert G. Crawford, Ph.D., M.B.A., M.S.I.S.
Department of Health Policy
Jefferson Medical College
1015 Walnut St., Ste. 115
Philadelphia, PA 19107

E-mail: albert.crawford@jefferson.edu