Healthcare costs in the U.S. are the highest in the world and are increasing rapidly. With increasing costs, employers and health insurance companies are trying to contain the cost of healthcare. This study aims at developing a predictive model relating enrollees’ healthcare insurance claims to their health risks while controlling for gender, age, and their previous year’s claims. Our study findings suggest that age, gender, and a previous year’s healthcare expenditure are strong predictors of healthcare cost and that none of the six biomarkers was a significant predictor. Even though these biomarkers failed to contribute to the predictive model, they are, nonetheless, important predictors of future chronic diseases, many of which are leading causes of death in the U.S.

**KEYWORDS** age, biomarkers, gender, healthcare cost, healthcare cost predictors

**INTRODUCTION**

Healthcare costs in the U.S. are the highest in the world (Smith, Cowan, Heffler, Catlin, and The National Health Accounts Team, 2006) and are increasing rapidly. In 2004, the U.S. spent $1.9 trillion on healthcare which accounted for 16% of its GDP. The per capita healthcare expenditure was $6,280. As a result, employers and health insurance companies are trying to contain the cost of healthcare (Feldstein, 2007). One of the best ways to contain costs is to keep people healthy (Edington, Yen, and Witting, 1997; Yen, McDonald, Hirschland, and Edington, 2003; Breslow, Fielding, Herrman, and Wilbur, 1990).
There have been many studies that investigate the relationship between improving health behaviors (exercise, diet, not smoking, etc.) and total healthcare expenditure (Edington et al., 1997; Yen, Schultz, Schnueringer, and Edington, 2006; Ozminkowski et al., 2006; Musich, Hook, Barnett, and Edington, 2003). Positive lifestyle changes are logically associated with a decrease in health risks. Specifically, these risks are associated with chronic diseases such as heart disease, stroke, diabetes, and some types of cancer which are the leading causes of death in the U.S. (Kung, Hoyert, Xu, and Murphy, 2007). However, these studies relied on participants' self-report of their health behaviors and often quantified total healthcare expenditures as an aggregate amount. Thus, it is difficult to determine who changed their lifestyle and decreased their healthcare expenditures.

This study utilizes a data set that links enrollees' health risks and their insurance claims with an individual enrollee as the unit of analysis. According to the medical literature, those diseases that are leading causes of death in the U.S. are highly related to health risks. These health risks include biomarkers such as body mass index (BMI), cholesterol levels (LDL, HDL, and total), fasting blood glucose level, and blood pressure (Marre, 2007; Hall, Crook, Jones, Wofford, and Dubbert, 2002; Haapaniemi, Hillbom, and Juvela, 1997; Wu et al., 2006; Calle, Rodriguez, Walker-Thurmond, and Thun, 2003). High values of BMI, cholesterol levels (LDL and total), fasting blood glucose level, and blood pressure are risk factors for developing cardiovascular diseases. Breast, esophageal, colon/rectum, and stomach cancers are related to higher BMI values. Higher glucose levels are related to diabetes and its associated health problems. This study investigates the role of these biomarkers as predictive factors of healthcare cost. It is also well known that a small number of people utilize healthcare more than others (Moeller, Cohen, Mathiowetz, and Wun, 2003; Cohen, Ezzati-Rice, and Yu, 2006) and that they may continuously utilize healthcare more (Diehr, Yanez, Ash, Hornbrook, and Lin, 1999). Thus, a previous year’s healthcare expenditure could be related to a subsequent year’s healthcare expenditure. This study includes the previous year's health insurance claim as an independent variable. Some demographic variables such as age and gender are also related to those chronic diseases. Generally, females utilize more healthcare than males and older people tend to utilize healthcare more than their younger counterparts (Cohen et al., 2006). Thus, those variables are also included in our model as control variables.

This study aims at developing a predictive model relating enrollees' healthcare insurance claims to their health risks while controlling for gender, age, and their previous year’s claims. The health risks in this study include BMI, cholesterol levels (LDL, HDL, and total), fasting blood glucose, and blood pressure. We hypothesize that people with high health risks would utilize healthcare more than people with low health risks. This study defines health risks based on the guidelines published by the
Department of Health and Human Services, the Centers for Disease Control and Prevention (CDC), and the National Institutes of Health (NIH) (Centers for Disease Control and Prevention, 2007; National Heart, Lung, and Blood Institute, 1998, 2006).

BMI is computed with height and bodyweight and is used to classify possible weight problems. It is calculated as follows: weight (kg)/[height (m)]² or weight (lb)/[height (in)]² × 703. The normal range of the BMI for adults is defined as 18.5–24.9. The overweight range is defined as 25.0–29.9 and the obesity range is 30 and over.

While a high level of blood cholesterol is a major risk factor for heart disease, it does not display symptoms. Actually, cholesterol is a necessary ingredient of the body, but too much cholesterol leads to narrowing of the arteries and, consequently, leads to heart disease. There are two types of cholesterol measures: low-density lipoproteins (LDL) and high-density lipoproteins (HDL). A high level of LDL is associated with a higher risk for coronary disease. On the other hand, a high level of HDL is associated with a lower risk because HDL reduces the level of cholesterol. A high level of total cholesterol is considered a greater risk. For LDL, the optimal level is lower than 130 mg/dL, the borderline level ranges between 130 mg/dL and 159 mg/dL, and the high risk level is 160 mg/dL and over. For HDL, the optimal level is 60 mg/dL and over, the low risk level ranges between 40 mg/dL and 59 mg/dL, and the high risk level is lower than 40 mg/dL. For total cholesterol, the desirable level is below 200 mg/dL, the borderline level ranges between 200 mg/dL and 239 mg/dL, and the high risk level is 240 mg/dL and over.

Glucose is a carbohydrate used by body’s cells as an energy source. It is an important substance for the human body, but a high level of glucose is associated with a high risk for diabetes. Diabetes affects more than 20 million Americans and about 54 million Americans have prediabetes. Most have adult onset diabetes referred to as Type II diabetes. Glucose levels up to 99 mg/dL are considered normal. Levels between 100 and 125 mg/dL are referred to as impaired fasting glucose or prediabetes. These levels are considered risk factors for Type II diabetes and its complications. The levels of 126 mg/dL and higher are associated with high risk and people with this level of blood glucose would consequently be diagnosed as having diabetes.

Like high blood cholesterol levels, high blood pressure (or hypertension) usually does not display any symptoms and patients normally do not realize that they have high blood pressure. Having high blood pressure increases one’s risk for developing heart disease, having a stroke, and developing other serious cardiovascular diseases. It is estimated that nearly one out of three American adults has high blood pressure. The blood pressure readings are measured in millimeters of mercury (mmHg) and usually given as two numbers. The first number is the systolic blood
pressure reading (contraction pressure). The second number is the diastolic blood pressure reading (sustained pressure between contractions). It is noted that blood pressure readings continually change. Normal blood pressure is a systolic pressure of less than 120 mmHg and a diastolic pressure of less than 80 mmHg. Pre-hypertension is defined as a systolic pressure of 120–139 mmHg or a diastolic pressure of 80–89 mmHg. Persons with prehypertension are at increased risk to progress to hypertension. High blood pressure or hypertension for adults is defined as a systolic pressure of 140 mmHg and higher or a diastolic pressure of 90 mmHg and higher. Within the hypertension group there are two levels; Stage 1 and Stage 2. The Stage 1 is defined as a systolic blood pressure of 140–159 mmHg and a diastolic blood pressure of 90–99 mmHg. Stage 2 is defined as a systolic blood pressure of 160 mmHg or over and a diastolic blood pressure of 100 mmHg or over. If the systolic and diastolic blood pressure levels are in different categories, blood pressure status is defined according to the higher category.

METHODOLOGY

This study measures enrollees' biomarkers and health claim costs over two years for a data set of 2,103 enrollees from five employers: one manufacturing company with 1,000 employees in Michigan, one county government with 400 employees in Arkansas, one religious organization with 250 employees in Tennessee, one county government with 225 employees in Indiana, and one bank with 40 employees in Indiana. These enrollees were tested between June and November in 2006 at a test site, a rented facility, or onsite. The data include variables such as BMI, cholesterol levels, glucose levels, blood pressure, gender, age, and the previous year's healthcare expenditures. The dependent variable is the 2006 health claim that includes paid claim and prescription drugs. It does not include out-of-pocket expense. The cases with accidental expenditures are excluded from the analysis.

This study recodes the health risk levels defined by the NIH and the CDC in the U.S. Department of Health and Human Services. The health risk levels are recoded in a way that larger numbers indicate higher levels of risk. For the BMI, 0 is assigned if the values are between 18.5 and 24.9. When the values are between 25.0 and 29.9, 1 is assigned. When the values are 30 and over, 2 is assigned. The values below 18.5 are considered underweight and although they may show a health risk, it would be a different type from current leading causes of deaths. Thus, they would need to be analyzed differently and are, therefore, excluded from this analysis.

For LDL, 0 is assigned when values are lower than 130 mg/dL. When values are between 130 mg/dL and 159 mg/dL, 1 is assigned. When values
are 160 mg/dL and over, 2 is assigned. For HDL, 0 is assigned when values are 60 mg/dL and over. When values are between 40 mg/dL and 59 mg/dL, 1 is assigned. When values are lower than 40 mg/dL, 2 is assigned. Recall that a higher level of HDL is considered a lower risk. For Total cholesterol, 0 is assigned when values are below 200 mg/dL. When values are between 200 mg/dL and 239 mg/dL, 1 is assigned. When values are 240 mg/dL and over, 2 is assigned.

For glucose, 0 is assigned when values are up to 99 mg/dL. When values are between 100 and 125 mg/dL, 1 is assigned. When values are 126 mg/dL and higher, 2 is assigned.

For blood pressure, 0 is assigned when a case shows both a systolic blood pressure of less than 120 mmHg and a diastolic blood pressure of less than 80 mmHg. When a case shows a systolic blood pressure of 120–139 mmHg or a diastolic blood pressure of 80–89 mmHg, 1 is assigned. When a case shows a systolic blood pressure of 140–159 mmHg or a diastolic blood pressure of 90–99 mmHg, 2 is assigned (Stage 1 category). When a case shows a systolic blood pressure of 160 mmHg or higher or a diastolic blood pressure of 100 mmHg or higher, 3 is assigned (Stage 2 category). If the systolic and diastolic blood pressure levels are in different risk categories, blood pressure risk is defined according to the higher level.

The healthcare expenditures of 2005 are used to define the previous year’s healthcare expenditure. The expenditures are dichotomously divided into two groups: high expenditure and low expenditure by using the median of $702.12 as the break point. Because it is not possible to distinguish between people who left their employment and people who did not incur any healthcare expenditure in this data set, those cases with a value of 0 are excluded from the analysis.

The control variables used in this study are gender and age. The gender is recoded as female or not-female. The age variable is recoded as old age (40 and over) and young age (younger than 40) groupings.

The dependent variable is the health claim of 2006. The distribution of healthcare claims is not normally distributed and is usually highly positively skewed with many enrollees spending very little and a small number of enrollees spending a large amount (Diehr et al., 1999). This distribution is not suitable to a multiple linear regression analysis. Thus, this study utilizes a multiple logistic regression analysis. Cases are divided into two groups: high cost users and low cost users. There are several possible cutoff points according to the literature (Yen et al., 2006; Moeller et al., 2003; Cohen et al., 2006). The high user group in this study is defined as those above the median (50th percentile) costs. The median of the 2006 healthcare expenditure was $501.96. The data were analyzed with the objective of developing the best model to predict the set of enrollees who would potentially utilize healthcare as a high user.
RESULTS

Of the 2,103 cases in the provided data set, 827 are complete in terms of the analysis variables, and these variables are used to create our predictive model. There are 472 males and 355 females. The age distribution ranges from 18 to 90 and with the dichotomized age variable, there are 316 cases under 40 and 511 cases in the older age group. The healthcare claims for 2006 range from $3.60 to $185,014.07. The mean cost of the health claims is $2,780.76 and the standard deviation is $9,528.20. As noted above, the median cost was $501.96. Thus, the distribution of the health claim is very positively skewed. This distribution again indicates the appropriateness of using multiple logistic regression analysis to predict a dichotomous high-low cost dependent variable. It should be noted that the mean cost of this study is much smaller than that of the national average. That is because the costs in this study only include paid claims and prescription drug expenditures and these enrollees are generally healthier than the general population.

The distributions of risk levels for total cholesterol, LDL, HDL, BMI, glucose, and blood pressure are shown in Tables 1, 2, and 3. The higher number in the risk levels indicate a higher risk of developing chronic diseases and consequently of incurring higher healthcare expenditures.

For total cholesterol, 61% of the people have a desirable level, 30% have a borderline level, and 9% have a high risk level. For LDL, 78% have an optimal level, 17% have a borderline level, and 5% have a high risk level. For HDL, 39% have a protective level, 50% have a borderline level, and 11% have a major risk level. Thus, regarding the cholesterol levels, almost one in four people needs to control LDL, a bad cholesterol, and more than half of the people need to increase HDL, a good cholesterol.

For BMI, 30% show a normal weight level, 38% show an overweight level, and 32% show an obesity level. For glucose level, 64% show a normal level, 32% show an impaired glucose level, and 4% show a high risk of diabetes. The overweight issue has been receiving much attention in the U.S. and its increasing prevalence may be related in part to the increasing consumption of foods, such as oversized servings (Young and Nestle,

<table>
<thead>
<tr>
<th>TABLE 1</th>
<th>Distributions of Risk Levels for Total Cholesterol, LDL, and HDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk levels</td>
<td>Total cholesterol Frequency</td>
</tr>
<tr>
<td>0</td>
<td>507</td>
</tr>
<tr>
<td>1</td>
<td>246</td>
</tr>
<tr>
<td>2</td>
<td>74</td>
</tr>
<tr>
<td>Total</td>
<td>827</td>
</tr>
</tbody>
</table>
This population clearly reflects the high incidence of overweight Americans. Only one in three people maintains normal weight. Although almost two out of three people currently show normal glucose levels, they may be at risk of increasing their glucose levels in the future because of being overweight.

For blood pressure, 28% show a normal level, 60% show a prehypertension level, 9% shows the Stage 1 hypertension level, and 2.7% shows the Stage 2 hypertension level. Although about 10% of people show high blood pressure levels, two-thirds of people have prehypertension levels. With these biomarkers, two-thirds of people show a normal level with total cholesterol, LDL, and glucose. However, only one-third of people show a normal level with HDL, BMI, and blood pressure.

A logistic regression analysis was performed to measure the predictability of the cost of healthcare in 2006 based on the six biomarkers taken as ordinal covariates, gender (female or not), age (above or below 40), and the previous year's costs (above or below the median). A stepwise procedure was applied utilizing a standard 0.05 entry criterion. The study population consisted of 827 individuals with all six recoded biomarkers and healthcare costs for 2005 and 2006. As previously discussed the break point separating low and high costs for 2006 was defined by the median ($501.96). Similarly, healthcare costs for 2005 were defined as high or low based on that year's median cost ($702.12). The result of multiple logistic regression analysis with stepwise procedure is shown in Table 4.

The logistic regression analysis with stepwise procedure selected the previous year's (2005) healthcare cost, gender, and age as significant

### TABLE 2 Distributions of Risk Levels for BMI and Glucose

<table>
<thead>
<tr>
<th>Risk levels</th>
<th>Frequency</th>
<th>Percent</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>247</td>
<td>29.87</td>
<td>529</td>
<td>63.97</td>
</tr>
<tr>
<td>1</td>
<td>313</td>
<td>37.85</td>
<td>264</td>
<td>31.92</td>
</tr>
<tr>
<td>2</td>
<td>267</td>
<td>32.29</td>
<td>34</td>
<td>4.11</td>
</tr>
<tr>
<td>Total</td>
<td>827</td>
<td>100.00</td>
<td>827</td>
<td>100.00</td>
</tr>
</tbody>
</table>

### TABLE 3 Distributions of Risk Levels for Blood Pressure

<table>
<thead>
<tr>
<th>Blood pressure</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>232</td>
<td>28.05</td>
</tr>
<tr>
<td>1</td>
<td>497</td>
<td>60.10</td>
</tr>
<tr>
<td>2</td>
<td>76</td>
<td>9.19</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>2.66</td>
</tr>
<tr>
<td>Total</td>
<td>827</td>
<td>100.00</td>
</tr>
</tbody>
</table>
predictors of the healthcare cost of 2006. Because the concept of $R^2$ is not appropriate with logistic regression analysis, the Hosmer and Lemeshow test of goodness of fit was conducted. This test compares the predicted probabilities (e.g., the probability that a female, over 40, that spent above the 2005 median will spend above the median in 2006) to the relative frequencies of the observed dependent variable (i.e., being above or below the 2006 median) of groupings defined by percentile breaks of the calculated probabilities. The measurement was 7.2742 ($p = .2962; \text{df} = 6$) and thus, randomness can be rejected in favor of accepting the model as a reasonable fit. All biomarkers (total cholesterol, LDL, HDL, BMI, glucose, and blood pressure) were not selected by the model. We did additional logistic regression analysis forcing the inclusion of all independent variables. The result confirmed that all six biomarkers were not statistically significant at $\alpha = 0.05$ (not shown here).

The odds for enrollees in the high level ($702.12$ or over) of healthcare expenditures in the previous year were 2.7 times more likely than for the enrollees in the lower level healthcare expenditures to spend more than $501.96$ in 2006, holding everything else constant (OR = 2.674; 95\% confidence interval [CI] = 2.008, 3.561). The odds for women were 1.7 times more likely than for men to spend more than $501.96$, holding everything else constant (OR = 1.68; 95\% [CI] = 1.258, 2.242). The odds for enrollees 40 years and over were 1.4 times more likely to be higher spenders than for those younger than 40 years, holding everything constant (OR = 1.449; 95\% [CI] = 1.079, 1.946). Thus, enrollees spending more than the median level on healthcare in the previous year, being female, and being 40 or older are strong predictors of higher healthcare cost in the next year.

**DISCUSSION**

The results indicate a clear relationship between current healthcare expenditure and previous year’s healthcare expenditure, gender, and age. It is noteworthy that none of the biomarkers demonstrated a strong relationship with the current healthcare expenditure. This result is counterintuitive from many previous studies and practices. When we analyzed only biomarkers,
age, and gender, some biomarkers showed statistical significance. Thus, they would have remained in the predictive model. However, when the previous year’s expenditure was included in the model, they all lost their statistical significance. The previous year’s expenditure was so strong that no biomarkers were predictively comparable. Most of the previous studies have investigated the relationship between healthcare expenditure and behavioral changes or biomarkers without a previous year’s healthcare expenditure variable. We would have concluded in the same way, if we had not included the previous year’s healthcare expenditure in the model. Our study has found the more powerful predictor to predict the next year’s healthcare expenditure. Thus, our study result is not an isolated finding. One article claims that previous utilization is a strong predictor of future utilization (Diehr et al., 1999), and another study argues that there was no relationship between healthcare cost and participation to health promotion (Sciacca, Seehafer, Reed, and Mulvaney, 1993). When considering the nature of chronic diseases, it is not conceivable that we can find a clear relationship between healthcare cost and biomarkers or behavior changes in a short period of time. Many people with high risk do not know that they are at high risk. They do not show any symptoms and thus, they do not have reasons to utilize healthcare. However, they may eventually succumb to chronic diseases and incur high cost in the future. It is also noted that when an insurance company assesses a health risk of a group of employees, they usually include age, gender, and past years’ healthcare expenditures to determine health insurance premium (Feldstein, 2007). Our study confirms that older people tend to utilize healthcare more than younger people. Older people have more chronic diseases. Women utilize healthcare more than men because of reproductive issues. People who utilize healthcare more often tend to utilize it in subsequent years. Many insurance companies use experience ratings when they determine health insurance premiums. It is not possible for them to obtain prospective customers’ biomarkers before they insure them, but those companies can define a financial risk based on age, gender, and healthcare utilization experience. This study confirms that this practice is possible as long as it is a short period of time.

CONCLUSION

Our findings suggest that age, gender, and a previous year’s healthcare expenditure are strong predictors of healthcare cost and that none of the six biomarkers was a significant predictor over a short period. One might infer that biomarkers are not important to control healthcare cost, and may further assume that it is not important to pay attention to biomarkers. What we have found, however, is that biomarkers are not strong predictors of the next year’s healthcare expenditure. When an insurance company assumes a
financial risk, it would not insure a group of people for just one year. The insurance company would insure those people continuously for many years. Thus, even though these biomarkers failed to stay in the predictive model, they are, nonetheless, important predictors of future chronic diseases, many of which are leading causes of death in the U.S. It takes time for those chronic diseases to develop, to show symptoms, and to require medical attention. It is well known that a small portion of people utilize most of healthcare resources and that patients utilize healthcare the most in their last year of life. We want to avoid premature death and push back this last moment as much as possible. Therefore, for a short term, the previous year’s healthcare expenditure is a strong predictor for the next year’s healthcare expenditure. However, biomarkers would probably be strong predictors of long term healthcare expenditures. Healthcare marketing managers are well advised to pay attention to their objective, especially if it is based on a short-term or a long-term strategy. It is also recommended for researchers that future studies utilize a longitudinal design and follow those participants for many years. When these studies include many relevant biomarkers as well as sociodemographic variables of the participants, they would reveal who actually develop these targeted chronic diseases and who will need more or less costly medical services over their lifetime.

REFERENCES


