Katherine M Kovacs
December 2013

THE EFFECT OF LIFESTYLE FACTORS ON US HEALTHCARE EXPENDITURES AMONG ADULT DIABETICS USING 2011 MEDICAL EXPENDITURE DATA

Abstract: *A retrospective, cross-sectional analysis of 2011 Medical Expenditures data was conducted to assess the impact of health behaviors/ controllable health factors on the total healthcare expenditures of adult diabetic patients in the US. Main effects and interaction effects were explored using multiple linear regression models. A linear association between mean log expenditures and exercise, diet modification and smoking was indentified.*

The study suggests insured diabetics have lower healthcare expenditures if they engage in moderate to vigorous exercise at least 3x/week. Because insured diabetics represent over 90% of the diabetic population, exercise has the potential to both reduce health care expenditures and improve health outcomes. Additional research is necessary given the limitations of self-reported data.

INTRODUCTION:

Hypothesis: Diabetes is one of the most critical health and economic issues facing the US today. According to a review published December 3, 2013 in *The Lancet Diabetes & Endicronology*, there has been an "inexorable and unsustainable increase in global health expenditures attributed to diabetes". Prevalence has increased due to high and growing rates of obesity and the sedentary lifestyle of Americans. The American Diabetes Association estimates that 25.8 million people in the US (or 8.3% of the population) have diabetes, of which 18.8 million are diagnosed. It is the leading cause of kidney failure, nontraumatic lower-limb amputations, and new cases of blindness among US adults. It is a major cause of heart disease and stroke and is the seventh leading cause of death in the Unites States.[1]

Diabetes is also a major driver of health care expenditures in the United States. According to the 2011 National Diabetes Fact Sheet, the total cost of diagnosed diabetes in 2012 is $245 billion, including $176 billion in direct medical costs. [2] The largest proportion of medical cost of those with diabetes was incurred by those over the age of 65. [3] The largest components of direct medical costs were for hospital inpatient care and nursing home care, followed by pharmacy and supplies and outpatient supplies. [4] Because diabetes is a significant driver of medical costs and morbidity in the US, a better understanding

of how lifestyle factors are associated with expenditures could help inform and fund public health outreach to the diabetic population.

A number of studies have been conducted in this area, but have not been strongly conclusive regarding the effect of diet and exercise on health care expenditures. For example, a 2011 study conducted by the Health Economics Group in the UK among diabetes concluded that lifestyle interventions to reduced T2DM were not considered cost-effective, although there was a large degree of uncertainty surrounding these estimates.[5] Additional studies are needed to better understand how to lifestyle changes may enhance diabetes outcomes while reducing expenditures, a goal consistent with US health care reform efforts.

I. METHODS:

Data and Variables: This is a retrospective, cross-sectional study using dataset from the Medical Expenditure Panel Survey (MEPS), a national probability panel survey co-sponsored by the Agency for Healthcare Research and Quality (AHRQ). The 2011 dataset was utilized, which was published in September 2013 and is the most recent data available. MEPS is designed to provide nationally representative estimates of health care use, expenditures, sources of payment and insurance coverage. The survey collects data on 35,000 individuals sampled from the non-institutionalized civilian population of the US.[6] The sample design is a multistage area probability design with disproportionate sampling to facilitate the selection of oversamples of population of interest. The person weights were applied in the exploratory analysis and model building procedures to provide national estimates.

The study subjects were restricted to adults aged 18 years or older who self-reported a diagnosis of diabetes. The unit of analysis was one individual patient. The survey questions do not delineate between Type 1 and Type 2 diabetes; therefore, both types of diabetes are included. The 2011 dataset 2,590

diagnosed adult diabetics (i.e. unweighted sample size), representing a weighted sample of 22.7 million diabetics in the US population. All analyses were conducted using SAS 9.3 (SAS Institute, Cary, NC). Dummy variables were created for all categorical variables.

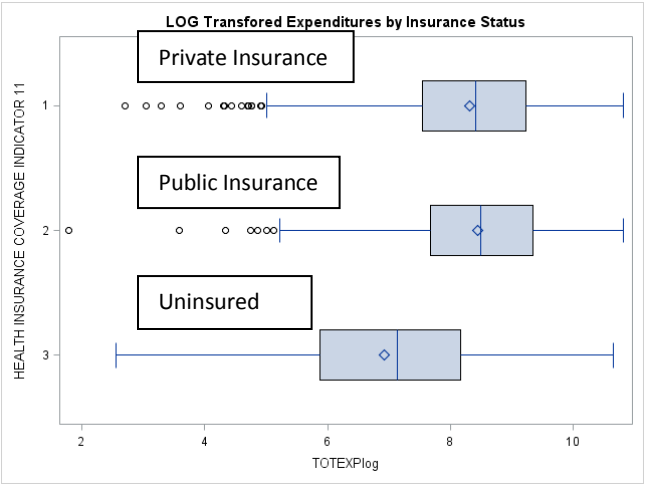
Hypothesis: To add to the body of evidence regarding the impact of interventions on diabetes expenditures, this analysis is designed to determine if a linear association exist between specific lifestyle factors (specifically, current smoking status, engaging in moderate-vigorous exercise >= 3x/week, diet modification as a treatment for diabetes, and BMI) against mean total health care expenditures (i.e. the response variable) among adult diabetics 18 years of age and older in the US. It should be noted that the health care expenditures in the study are based on total expenditures from all sources, i.e. include both governmental and individual expenditures.

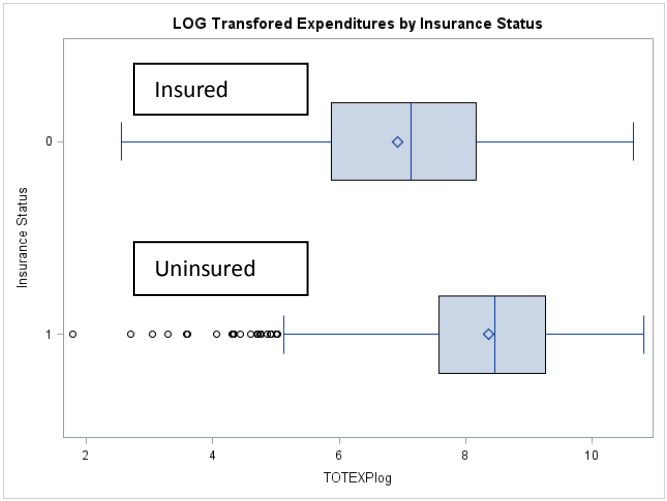
Age, gender, health insurance coverage, and personal income were explored as control variables. Interaction effects between variable were also assessed. The null and alternative hypotheses are:

H_0: Y (Mean medical expenditures) = B_0 + E *(after controlling for age, gender, health insurance coverage, and personal income)*

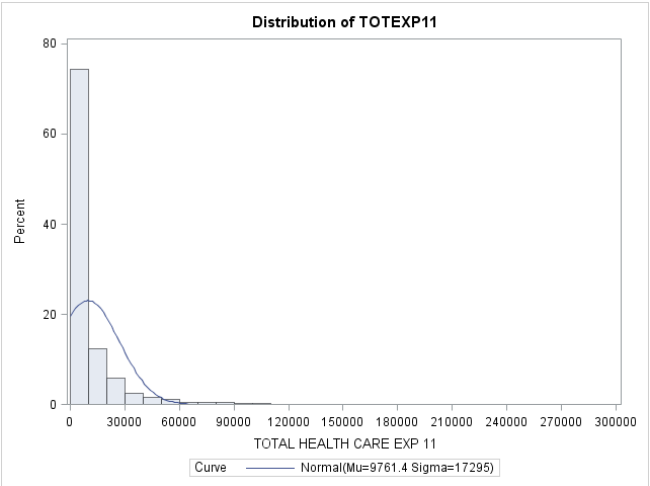
H_a: Y (Mean medical expenditures) = B_0 + B_1 Smoking + B_2 BMI+ + B_3 Exercise + B_4 DietModification + C1 Age + C2 Gender+ C3 Insurance Status + C4 PersonalIncome + E (*plus an examination of possible interaction effects*), where at least one $B_j \neq 0$

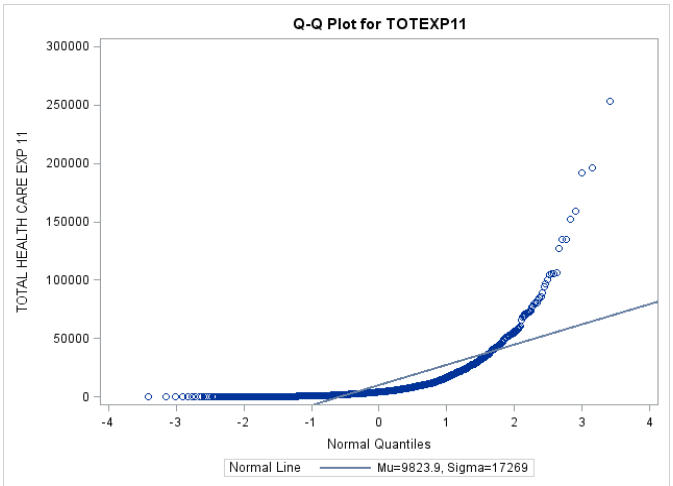
EXPLORATORY ANALYSIS: The initial step in the exploratory analysis was to visually examine the dataset. All observations without complete information were eliminated from the dataset. This step reduced the total number of observations from 2,590 to 1,977 observations. All categorical variables were dummy coded for further analysis. The Insurance variable was a originally a 3-level variable that was recoded to a 2-level variable due to the similarity between private and public health insurance on mean health expenditures and to enhance analysis and interpretation.





An inspection of the response variable, total health care expenditures, was highly skewed with extreme values at both tails, which violated the normality assumption, as shown below. A Shapiro-Wilks test statistic was not available for the original dataset because the number of observations exceeded 2,000.





Various transformations were employed including square root, reciprocal, and log. Further refinement of the dataset was conducted to include only expenditures greater than 0 and less than $50,000, resulting in a final sample size of 1,681 unweighted observations, corresponding to 16 million adult diabetics or over 70% of the diabetic population. None of the transformations were successful in achieving true normality as per Shapiro-Wilks testing and examination of QQ plots; therefore, a Box-Cox transformation was performed to inform the transformation, which suggested a lambda of .16. Exponentiating the response variable by .16 was successful in improving the normality as per a visual inspection of the distribution and QQ-plot, but the Shapiro-Wilks test statistic of .99 remained statistically significant (i.e. reject the null hypothesis of normality) due to the long tails of the distribution. Because a lambda between -.25 and +.025 is generally indicative of a log transformation, the decision was made to employ a log transformation to enhance interpretation of the models. In addition, the models were run with both transformations and there was little difference in the variable selection and residuals based on a lambda=.16 vs log transformation.

After the transformation, the relationship between the log of health care expenditures (i.e. the response variable) with the predictor and demographic variables was examined to assess the form of the relationship with the response variable, i.e. to better understand if the relationship is consistent with the linearity assumption that underlies multiple linear regression. The three (3) continuous, independent variables (i.e. BMI, age, total personal income) were explored using scatterplots and Pearson correlation coefficient matrix. A transformation for BMI was explored to see if the linear association between healthcare expenditures could be strengthened by a log and a square-root transformation, but neither transformation was highly successful and BMI was entered into the model in original units. Simple linear regression was run on each of the four predictor variables (i.e. BMI, smoking status, diet modification and exercise level) and the demographic variables to understand the individual relationships with the response variable and to assess for potential confounding, i.e. to assess if the co-variates were independently related to the response variable and the

predictor variable – the two conditions required to qualify as a confounding variable. The association between the dichotomous variables (i.e. the dummy variables created from the 2—category categorical variables, including exercise, insurance status, and all other categorical predictors) were tested using a chi-squared test . The association between categorical and a continuous variables were tested by t-test because the categorical variable has two categories.

All possible 2-way interaction s (i.e. age*BMI, diet modification*smoking, exercise*insurance etc) were checked via visual inspection (for interaction terms with a continuous variable) and via PROC GLM.

MODEL BUILDING: MAIN EFFECTS AND INTERACTION EFFECTS After performing the explanatory analysis, variable selection was assessed for a preliminary main effects model using the all-possible-regressions macro from SAS (ALLSUBSREG) which calculates R-squared, adjusted R-squared, Mallow's CP, predicted residual sum of squares , root mean square error , mean square error, in addition to other statistics. The criteria for model evaluation included all above statistics. An all-possible regression model was also run on the main effects (i.e. all independent variables without the interaction terms). The person weights were applied at the regression level to provide results that can be extrapolated to the adult diabetic population.

Two sets of hypotheses were tested. The first hypothesis, described earlier, was to assess if any of four health behaviors/health status factors were linearly associated with the mean log of total health care expenditures, after adjusting for demographic variables, i.e. a main effects model. After identifying an main effect model, an additional hypothesis test was conducted to ascertain whether a larger model including interaction terms provided a better explanatory model. The null and alternative hypotheses for reduced vs the full model are provided on the page below.

H_0: REDUCED MODEL (nested in full model): Y (Mean medical expenditures) = B_1 BMI + B_2 Insurance Status + B3 Exercise Level + B_4 Age *+E (i.e. b5=b6=b7=b8=b9=b10=b11=0)*

H_a: FULL MODEL (including interaction terms): Y (Mean medical expenditures) = bo + B1 BMI + B2 Insurance State + B3 Exercise + B4 Age + B5 DietModification + B6 Male + B7 Smoker + B8 + Male*DietModification + B9 Exercise*Insured + B10 Age*Diet Modification + B11 Smoker*Age + e, , where at least one B_{5-} $B_{11} \neq 0$

Following exploration of the main effects model, interactions were assessed. The interaction model was explored based on findings in the exploratory analysis and evaluated based on maximized adjusted R2, minimized MSE, and a priori knowledge regarding the independent variable effects. The ALLSUBS macro was not used to evaluate the interaction model because the number of variables exceeded the maximum.

MODEL ADEQUACY: A global F-test conducted on a main effects and interaction effects model was highly significant (p<.0001). Variance inflation factors and chi-squared tests were used to assess multi-collinearity, which were not evident across any of the predictor variables as evidenced by the VIFs and Condition Index. Overall, the number of outliers relative to the large dataset (16/1681) was relatively small and further indication of good model fit and consistency with the normality assumptions.

MODEL ASSUMPTIONS: Existence and independence assumptions were assumed given stringent data collection efforts and sample size. The main effects and interaction models were assessed for the other multiple linear regression assumptions of linearity, normality, and homoscedastiticity using residual diagnostic tests and plots designed to detect violation of the modeling assumptions . Scatterplot of jackknife residuals and scatterplots of residuals vs each independent continuous variable were used to test the linearity assumption. The Jacknife plot was also used to assess the assumption of equal variance. Normality was assessed via scatterplots, Q-Q plots, and Shapiro' Wilks testing. Jacknife standardized residuals, leverage, Cook's distance, Dffits, DBetas were all examined to help identify outliers. Residual plots were examined to detect model lack of fit and unequal variances and added-variable plots were used to test assumptions of

linearity in the final model. Added Variable Plots were also examined for the continuous variables to assess the linearity assumption of multiple linear regression. The final model met all assumptions of MLR, with the aforementioned caveats regarding the normality assumption.

RESULTS:

Table 1) outlines the key variables associated with the study population. The mean age of an adult diabetic in this population was 61 years, with the majority between 50 and 70 years of age. A slight majority were male (56%) and less than 10% were uninsured. In terms of lifestyle factors, the majority (86%) were non-smokers and over three-quarters self-report using diet modification as a way to treat their diabetes. However, only a minority maintained a healthy weight and exercise routine. Only one-third engage in moderate to vigorous exercise at least 3 times per week. The average BMI in the sample is indicative of obesity (mean 32>30; 30 as the threshold of being obese).

Table 1) Descriptive Statistics of Study Population – Adults with Diabetes 18+

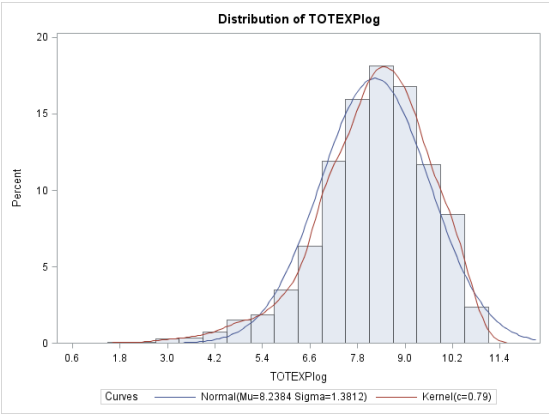
Variable Name	Mean	Standard Deviation	Range
Age	**60.5**	**13.7**	**18 – 85**
Gender (%)			
Female	46%		
Male	54%		
Total Personal Income	**$28,059**	**$27,430**	**$3 - 238,632**
Health Insurance Coverage (%)			
Private Insurance	55%		
Public Insurance	36%		
Uninsured	9%		
Treats Diabetes with Diet Modification(%)			
Yes	78%		
No	22%		
Currently Smoke (%)			
Smoke	14%		
Do not Smoke	86%		
Engages in Moderate/Vigorous Physical Exercise =>3X/week (%)			
Yes	35%		
No	65%		
BMI (Body Mass Index – imputed)	**32.05**	**7.40**	**16.6 - 68.3**
Total Health Care Expenditures *(from all sources including government and out-of-pocket payments)*	**$7,793**	**$9,249**	**$5 - $49,540**

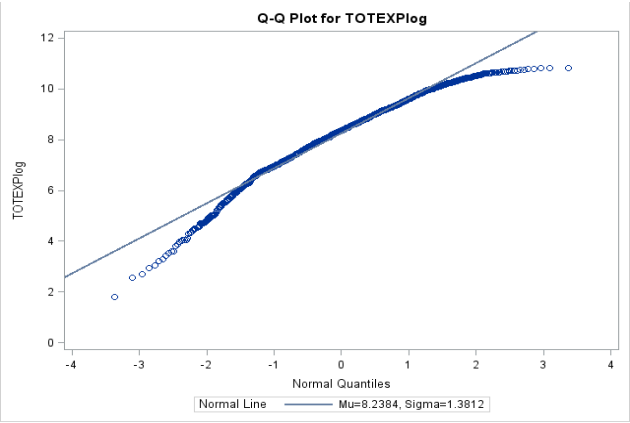
Demographic variables as covariates (based on prior research)

Health Behaviors/Lifestyle Factors as Predictor Variables

Response variable

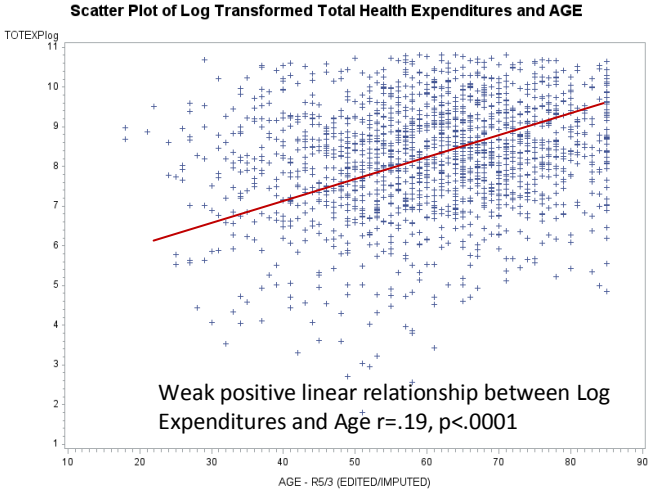
The response variable, Total Healthcare Expenditure for each individual adult diabetes patients, has a mean of $7,793 and a median of $4,303, suggestive of the highly right skewed distribution. There is also a significant standard deviation of $9,249. The expenditures range from $5 to $49,540 because the dataset was restricted to only those with expenditures greater than 0 and less than $50,000. This was necessary because the original distribution included an even wider range of expenditures (from 0 to $253,000) and it was not possible to achieve normality or near normality. A Box-cox analysis performed on the reduced set suggested a transformation of lambda = .16, and a log transformation was employed, as listed below.

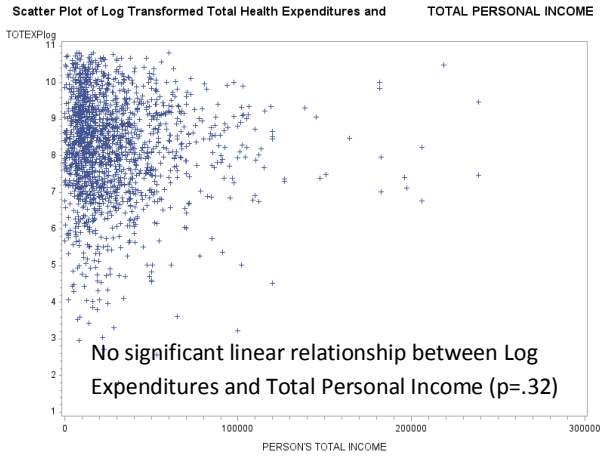


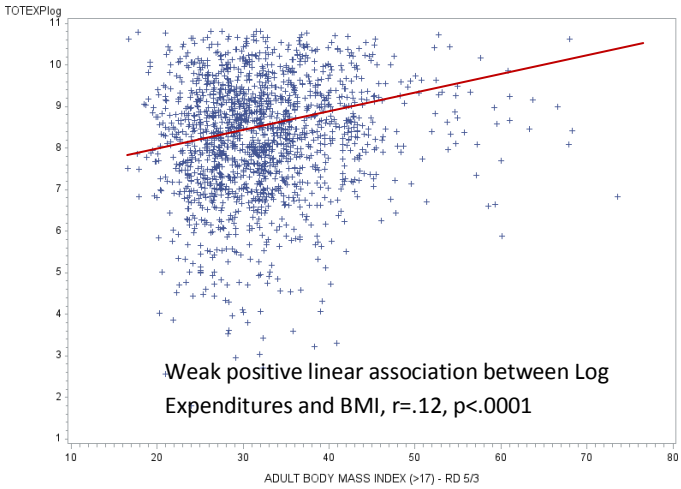


The log transformation on total health care expenditures significantly improved the distribution and QQ plot, as depicted in above. However, the Shapiro-Wilks test provided a test-statistic of .97 with a p-value of <.0001 suggestive of non-normality due to the extreme values that remain in the tails of the distribution. Still, the transformed distribution sufficiently meant the normality assumption to proceed with the analysis.

Only a subset of independent variables showed evidence of a linear association with the response variable in the exploratory phase. BMI and the log of Total Health Care Expenditures showed a statistically significant, but weak, positive linear association (r = 0.12, p-value of <.0001). Because the scatterplot of BMI and log expenditures showed some suggestion of a curvilinear relationship, the log and square root of BMI was examined vs log expenditures, but neither transformation significantly improved the association, and BMI was retained in original units. A weak positive association was also seen between the response variable and age (r=.19 with p-value <.0001). However, there was no evidence of a linear association between the response variable and total personal income (r=-.02 with p-value=.32, therefore, do not reject null of no linear association between the two variables). Scatter plots against the response variable are displayed below.

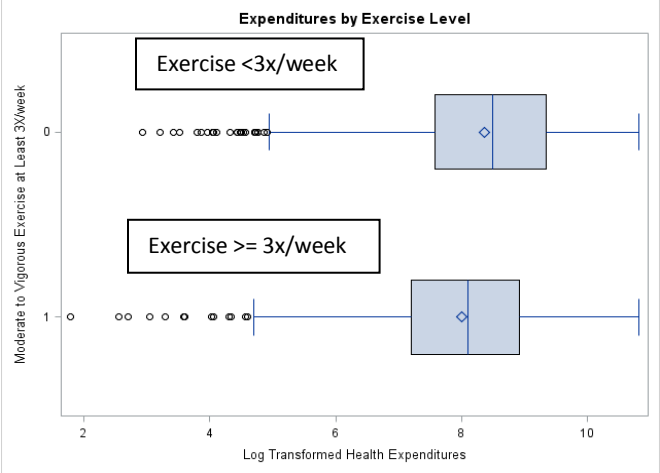




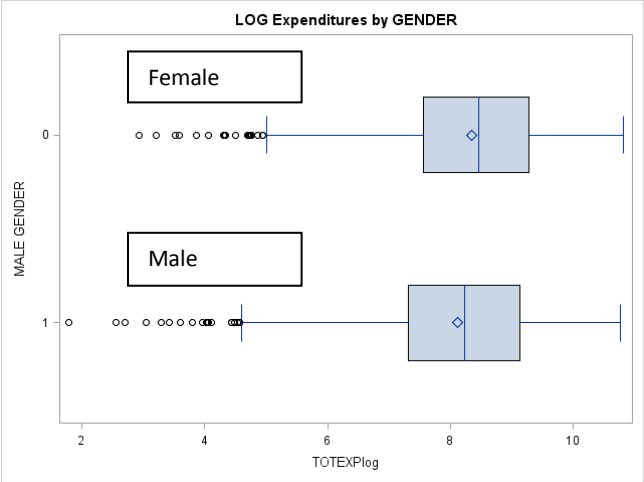


Various transformation were attempted to help linearize the relationship between personal income and the log expenditures (as shown above), but none were successful in achieving statistical significance.

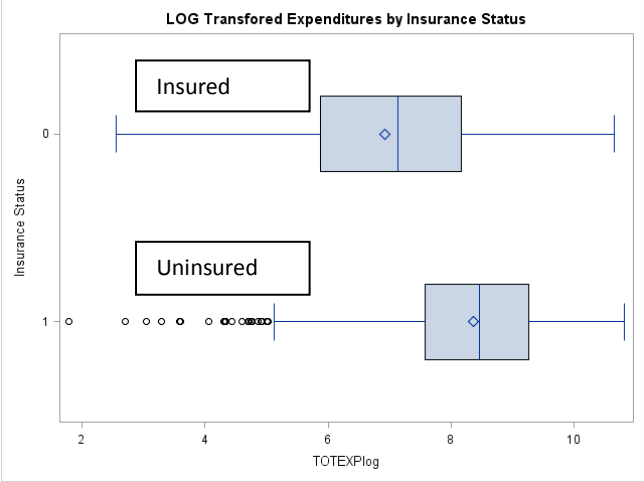
The relationship between the categorical predictor variables was then explored. Exercise level, gender, and insurance status showed statistical differences in mean log health care expenditures, while the other dichotomous variables did not show differences by log expenditures, including smoking – a surprising finding.



EXERCISE - A visual inspection of the box plot shows that mean log health expenditures are lower for those who exercise >=3x/week vs vs those who exercise < 3x/week. A t-test and SLR confirmed this finding (p<.0001). There appear to be more extreme values corresponding to lower expenditures in the exercise >= 3x/week group.



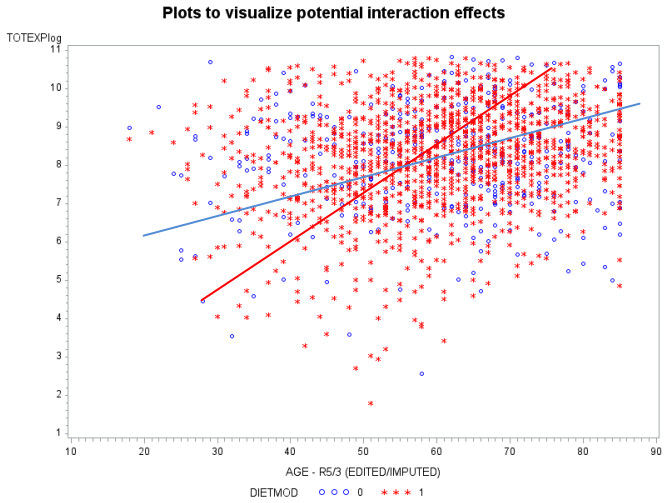
GENDER - T-test shows significant difference in mean log expenditures between males and females (p<.0005). Males have lower mean health care expenditures than females.



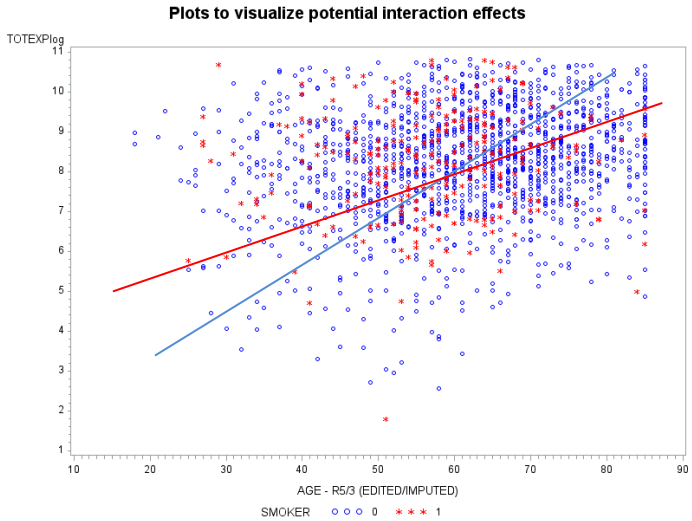
INSURANCE STATUS – The box-plot and T-test shows significant difference in mean log expenditures between the insured and uninsured groups (p<.0001), with greater variability of expenditures among the insured.

The relationship between the categorical variables with each other were also explored. A strong, positive association between diet modification and exercise was noted (chi-square statistic = 14.92, p =.0001). T-testing also showed a significant relationship between total income a significant negative linear association with insurance status (p <.0001).

All interaction terms were defined and explored. Plots to visualize interaction effects suggest possible interaction effects between age and smoking status, as well as age and diet modification, as shown below. A PROC GLM suggested additional interaction terms including gender and diet modification, and insurance status and exercise group.

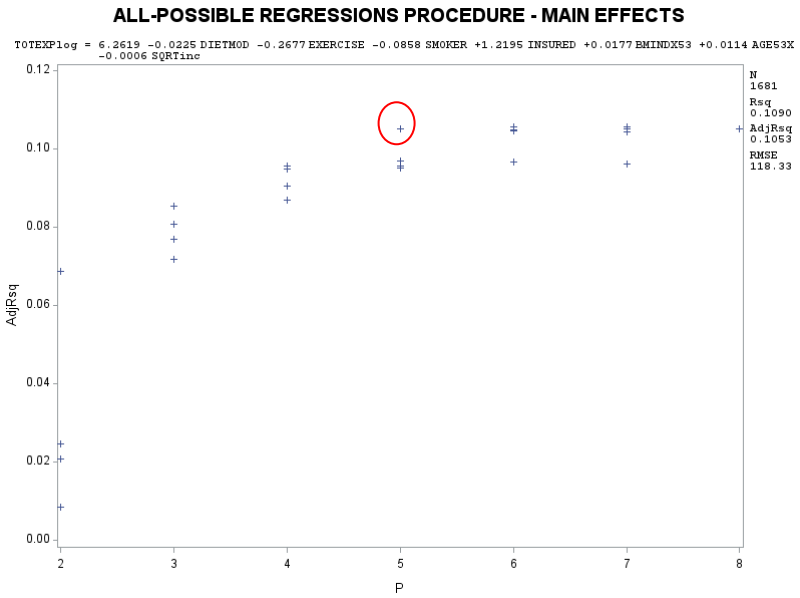


Plot suggests potentially different slope coefficients for age based on diet modification groups, ie. as age increases, mean log health expenditures increase at a higher rate among the diet modification group vs the non-diet modification group.



Plot suggest potentially different slope coefficients for age based on smoking status, i.e., as age increases, mean log health expenditures increase at a higher rate among the non-smoker group vs the smoker group.

A main effects model including all variables was constructed and assessed using the PRESS SAS macro and the all-possible regressions procedure.



Adjusted R2 = 10.5%

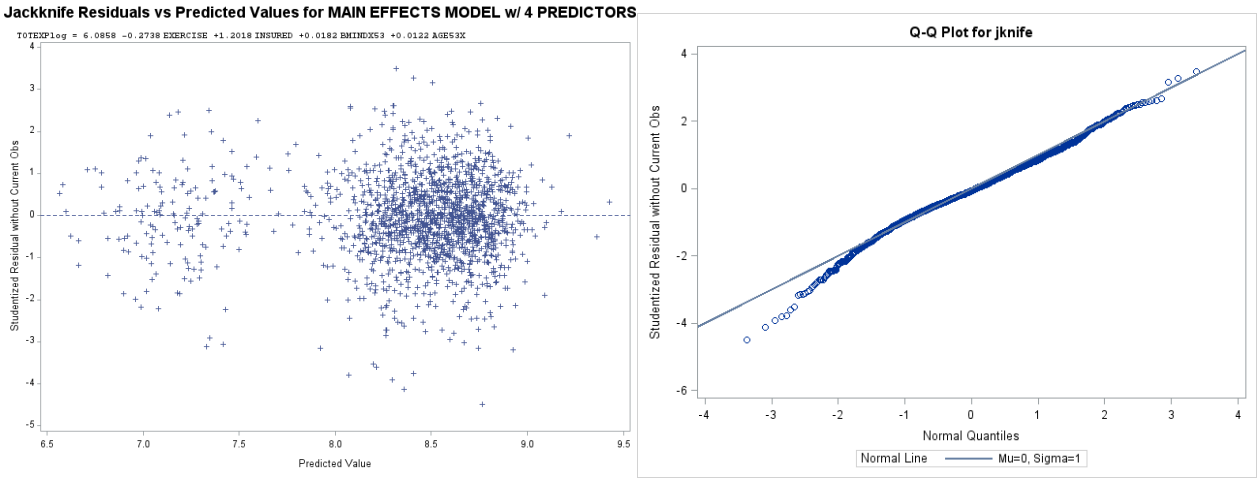
Mallow's C(P) = 5.37 (closes match to number of parameters = 6)

Press Statistic = 2801.4 (lowest among all possible models)

Both methods yielded the same model containing four predictors (or 5 parameters) including: Exercise level, insurance status, BMI, and Age. Model adequacy assessed by a Global F test was highly significant at alpha=.05 ($p<.0001$), i.e. reject the null hypothesis and conclude that at least one B1-4>0. All coefficients were also significant at $p<.0001$. However, the adjusted R2 was only 10.5%, suggesting that a number of key drivers of health care expenditures were missing from the model.

MAIN EFFCTS MODEL: Independent Variables	**Dependent variable - Log of Total Health Care Expenditures Adjusted R^2 = 10.5%**			
	Parameter Estimate	Standard Error	95% CI	p-value
Exercise >=3x/week	-0.27	0.0631	-.040 - .015	<.0001
Insured	1.20	0.1212	.960 – 1.44	<.0001
BMI	0.02	0.004	.010 - .026	<.0001
Age	0.01	0.0023	.008 - .017	<.0001
Intercept	6.08	0.234	5.63 – 6.54	<.001

The jackknife plot of the residuals vs the predicted demonstrated that the assumption of equal variance and normality were satisfied for the main effects model. A QQ-plot and the Shapiro-Wilks test suggested that near normality was achieved (i.e. S-W statistic = .989 p<.0001), consistent with the lack of perfect normality in the response variable.



In addition, there was no evidence of multicollinearity, with all Variance Inflation Factors between 1.002 and 1.038. In addition, the Condition Index was well below the threshold of 10.

INTERACTION MODEL

An interaction effect model was also explored based on the potential relationships identified in the exploratory phase. The all possible regressions model suggested an optimal model with 9 or 10 parameters.

Two interaction models were tested based on the adjusted R2, mallow C(P), and MSE. The two models had the following common variables: BMI, Gender, Diet Modification, Exercise Level, Smoking status, Insurance Status, Age, Gender*Diet modification, Exercise * Insurance Status, Age*Diet, and Smoker*Age. One model also included an interaction term for BMI*Age. The model without this term

was selected because the parameter estimate for BMI*Age was .0005 (the smallest of any of the predictor variables) and had a p-value on the border of significance at alpha = .05 (i.e. p=0.47).

FINAL MODEL – INTERACTION MODEL

Independent Variables	Dependent variable - Log of Total Health Care Expenditures Adjusted R^2 = 12.4%			
	Parameter Estimate	Standard Error	95% CI	p-value
BMI	.019	.0004	.011 to .023	0.0516
Male Gender*	.254	.131	-.0002 to .511	0.0123
Diet Modification	-.800	.320	-.143 to -.174	0.2101*
Exercise	.287	.223	-.162 to 0.735	0.0002
Current Smoker	1.52	.412	.710 to 2.320	<.0001
Insured (public or private)	1.47	.161	1.153 to 1.786	0.5431*
Age*	.003	.004	-.006 to .011	0.0047
Male Gender * Diet	-.414	.146	-.070 to -.127	0.0107
Exercise * Insurance	-.611	0.234	-1.07 to -0.141	0.0015
Age* Diet Modification	.016	.005	.006 to .026	.0002
Smoker*Age	-.028	.007	-.042 to -.014	<.0001
Intercept	6.32	.347	5.64 to 7.00	<.001

**Neither Gender nor Age are significant as main effect variables within the model, but are included because of significant interaction effects with both Gender and Age*

Fitted Model: Yi hat = 6.32 + .019 Male – 0.800 DietModification + .287 Exercise + 1.52 Smoker + 1.47 Insured + .003Age - 0.414 Male*DietModification -0.611 Exercise*Insurance +0.016 Age*DietModification – 0.028 Smoker*Age

The SAS statement, PCORR2, was used to compute the sum of squares for all independent variables in the model. Insurance status was shown to contribute most to the model (i.e. the highest type II sum of squares after controlling for all other variables) with a squared partial correlation Type II of .047, followed by BMI (.012), and the interaction term smoking * age (.009).

Interpretation of Parameter Estimates (for relevant variables):

B0 = When all variables are set to zero, mean log health care expenditures (expressed in dollars) is 6.32

B1 (BMI) = There is a .19 change in mean log health care expenditures for each unit increase in BMI, adjusted for all other variables.

B2 (Male Gender) = There is a .25 increase in mean log health expenditures (in dollars) if male gender vs female in the non-diet modification group, and adjusted for all other variables.

B4 (Exercise) = There is a 0.29 increase in mean log health expenditures (in dollars) if in the exercise >= 3x/week group vs the exercise <3x/week group among those who are uninsured, adjusted for all other variables.

B6 (Insured) = There is a 1.47 increase in mean log health expenditures (in dollars) in the insured group (either public or private) vs the uninsured group among those who exercise <3x/week, adjusted for all other variables.

B7 (Age)= For each one unit increase in age, there is a .003 increase in mean log health expenditures (in dollars) among the non-smoker and non-diet modification groups, adjusted for all other variables.

B8 (Male * Diet Modification) =The difference in mean log health expenditures (in dollars) between males and females in the diet modification group 0.60 lower than the difference in mean log health expenditures between males and females in the non-diet modification group, adjusted for all other variables.

B9 (Insured * Exercise) =The difference in mean log health expenditures (in dollars) between insured and uninsured in the exercise >=3x/week group is 0.60 lower than the difference in mean log health expenditures between the insured and uninsured in exercise <3x/week group, adjusted for all other variables.

B10 (Age*Diet Modification)=The change in mean log health care expenditures per year increase in age is .016 higher in the diet modification group vs the non-diet modification group, adjusted for all other variables.

B11 (Age*Smoking)=The change in mean log health care expenditures per year increase in age is .03 lower in the smoking group vs the non-smoking modification group, adjusted for all other variables.

The above interaction terms were interpreted by splitting the model. For example, the interpretation of B11 was derived as follows: Mean Log Health Expenditures = b0 +b5 Smoke+ b7 AGE +b11 Smoke*AGE + all others, where b11 = -0.3.
For SMOKING GROUP: (b0+b5) +(b7 + b11) AGE
For NON-SMOKING GROUP: bo + b7 AGE , therefore, b11 age i.e. the b11 coefficient represents the difference in the effect of age between the smoking and non-smoking groups.

An F test was conducted to determine whether or not the full (interaction) model provided a better explanation of the data than a reduced (main effects model, essentially a nested model of the interaction model), specifically testing if the additional terms have a s. The statistics are described below:

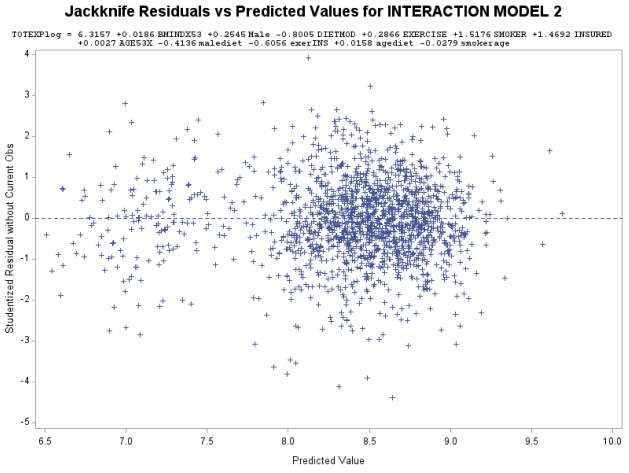
H_0: REDUCED MODEL (nested in full model): Y (Mean medical expenditures) = B_1 BMI + B_2 Insurance Status + B3 Exercise Level + B_4 Age *+E (i.e. b5=b6=b7=b8=b9=b10=b11=0)*

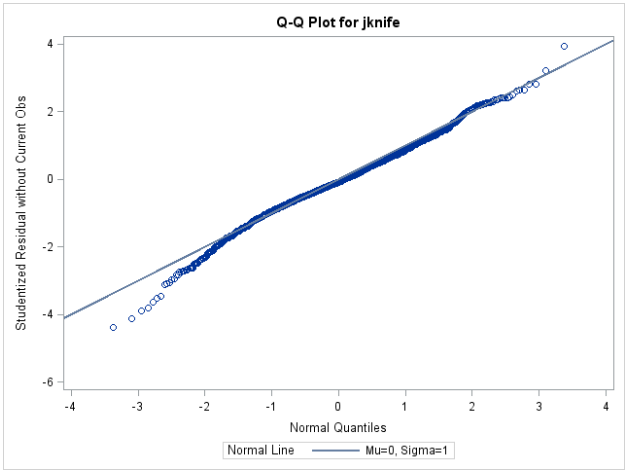
H_a: FULL MODEL (including interaction terms): Y (Mean medical expenditures) = bo + B1 BMI + B2 Insurance State + B3 Exercise + B4 Age + B5 DietModification + B6 Male + B7 Smoker + B8 Male*DietModification + B9 Exercise*Insured + B10 Age*Diet Modification + B11 Smoker*Age + e, , where at least one B_{5-} $B_{11} \neq 0$

F=((SSR (Full) – SSR (Reduced))/ (df F- df R))/MSE (Full) = (3,420,079*2,818,808/(11-4))/13703 = 6.27

Because 6.27 > X2 = F .05, 7, 1670 = 2.02, we reject the null hypothesis and conclude that at least one b5-b11 >0 .

The final, full model also met all model assumptions as per visual examination of the jackknife residual plot, which showed a random pattern similar to the residual plot from the main effects model. The Q-Q plot and Shapiro Wilks test (statistic = .99, p <.0001) were also suggestive of near-normality, but with evidence of extreme values at the tails of the distribution.



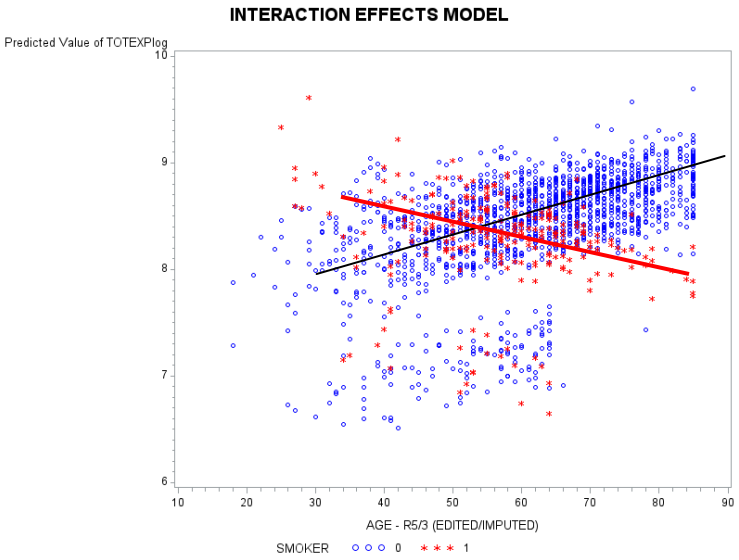


EXAMINATION OF POTENTIAL OUTLIERS: A total of 16 outliers were identified based on an examination of all diagnostics including Rstudent, hat diagnostics, covariance ratio, Cook's distance, and DFBetas. The outliers were not removed from the model but examined to ensure plausibility. The observations and the diagnostic values are listed in the table below.

Observation number	R Student	Hat Diagnostic	Cov. Ratio	Cook's Distance	DFBeta for BMI Index
536	-4.38	.024	.900	-.699	0.142
284	-4.11	.009	.900	-.3848	.0150
664	-3.90	.0057	.908	-.294	-.0386
49	-3.80	.00339	.912	-.338	.0873
999	-3.62	.007	.989	-.998	.1008
343	3.23	.0326	.966	.5921	.0321
138	3.93	.0262	.926	.645	-.209
885	2.31	.0478	1.018	.517	-.0147
67	2.01	.040	1.019	.412	.3816
610	2.06	.0604	1.040	.522	.1534
177	2.37	.0248	.990	.379	.1207
746	2.36	.0216	.980	.350	.1113
35	1.66	.1305	1.136	.644	.0947
1494	-3.53	.0021	.923	-.163	.0019
491	2.62	.0134	.972	.412	.3816
1112	-0.82	.0978	1.185	-.260	.0013

In addition, the hat diagnostics alone indicate a number of high leverage points. Using hii>2(k+1)/n (or .014 for this specific dataset) as a threshold, a total of 193 high leverage points were identified. These are an indication of an extreme value of xi (or a predictor variable). A review of the original observation shows these leverage points are driven by extremely high and low values of BMI; for example, one of the observations, #890, corresponds to a BMI of 53.8, indicating a morbidly obese individual. It is also possibly a coding error, although BMIs in this range are biologically plausible. Observation 35, another high leverage point, correspond to an underweight individual due to a BMI of 21.6. Overall, the number of outliers relative to the large dataset (16/1681) is relatively small and an indication of good model fit and consistency with the normality assumptions.

A plot of age by Smoking Status again predicted mean log expenditures displays the interaction effect.



II. DISCUSSION/CONCLUSION:

The full, interaction model provides evidence of a relationship between selected health behaviors and mean log health care expenditures in the US. Of particular significance is the finding that the difference in mean log health expenditures (in dollars) between insured and uninsured in the exercise >=3x/week group is 0.60 lower than the difference in mean log health expenditures between the insured and uninsured in exercise <3x/week group, adjusted for all other variables. Because insured diabetics represent over 90% of the diabetic population, vigorous-to moderate 3x/or more a week has the potential to both reduce health care expenditures and improve health outcomes among this population.

Another noteworthy finding is that age is an effect modifier of selected health behaviors on log mean health expenditures. This suggests that a one-size-fits-all approach might not be effective in reducing health expenditures as paradoxical effects were demonstrated. For example, the change in mean log health care expenditures per year increase in age is .03 **lower** in the smoking group vs the non-smoking modification group, adjusted for all other variables. This may be due to fewer office visits among smokers, who may be less likely to engage in positive health behaviors as compared to smokers. The model also suggests that having health insurance increases health care expenditures for diabetics, which has implications for funding of the ACA health reform.

The full model has limitations vs a main effects model. Although the interaction terms help to improve goodness of fit and provide useful information, the interpretation is more difficult than a simple main effects model with only BMI, Exercise level, Age and Insurance Status. From a public health perspective, the simpler main effects model is easier to interpret and may lend itself to broader dissemination among policymakers and other lay professionals.

A significant limitation of the modeling exercise was the skewed distribution and presence of many extreme values that required truncation of the dataset. Research from Bhattarat 2012 suggests that healthcare expenditure data often follows a Gamma, rather than a normal distribution.[7] The presence of extreme outliers not included in the analysis can influence the estimated savings or costs associated with health behaviors. Therefore, in an effort to normalize the distribution, it was necessary to restrict the dataset. As such, the findings cannot be extrapolated to the entire US adult diabetic population, but only to the subset with expenditures >0 and less than $50,000. Adult diabetics with expenses greater than $50,000 warrant further analysis, as this subgroup is a major driver of health care costs in the US.

Another limitation of both models is that the analysis relied on self-reported data. Because individuals tend to overestimate healthy behaviors and underestimate unhealthy behaviors, such as smoking, there is likely a bias toward the null and possible underestimation of effect sizes. In addition, previous studies have shown that MEPS tends to under-report total health care expenditures.

It is also noteworthy that the adjusted R2 in the full model is only 12.5% . Clearly, health care expenditures in diabetics is driven by numerous other factors not included in the model including total number of chronic health conditions, number and types of medications, compliance with medications, and frequency of health care visits, among other factors. However, while the R2 is low, the model is still useful in understanding the relationship between health behaviors and expenditures, i.e. it is useful for inference, but does not provide a useful model for predicting health care expenditures. Additional research is necessary given the above limitations, but provides additional evidence that exercise and a healthy BMI may result in cost savings.

Sources:

1 www.CDC.gov, American Diabetes Association

2 Evans JG et al, "A novel endocrinology-based wellness program to reduce medication expenditures and improve glycemic outcomes." Diabetes Metab Syndr. 2013 Apr-Jun;7(2):87-90. doi: 10.1016/j.dsx.2013.02.016. Epub 2013 Mar 22.

3 Herman H et al, "The Economics of Diabetes Prevention" Med Clin North America. 2011 March; 95(2):373-viii. doi: 10.1016/j.mena2010.11.010.

4 IBID.

5 Irvine L et al, "Cost-effectiveness of a lifestyle intervention in Preventing type 2 diabets." Int J Technol Assess Health Care. 2011 Oct; 27(4):275-82-. doi: 10.1017/S0266462311000365.

6 Lin, H-C et al, "Medication use and associated health care outcomes and costs for patients with psoriasis in the United States." Journal of Dermatological Treatment 2012; 23:196-202

7 Bhattari, G R, "Understanding the Outlier in Healthcare Expenditure Data." NESUG Proceeding, 2013, http://www.nesug.org/Proceedings/nesug13/116_Final_Paper.pdf