Vol. 12 (2026): Proceedings of the 2026 International Conference on Life Sciences and Health Engineering (IC-LSHE 2026)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/lshe.20260111

Determinants of U.S. Health Insurance Charges: Evidence from Multivariate Regression

Pengyuan Qian*

School of Mathematics and Statistics, Ningbo University, Ningbo, 315211, China

**Corresponding author: Pengyuan Qian*

Abstract

U.S. national health expenditures accounted for 17.3% of GDP, raising persistent concerns among insurers, policymakers, and households. Using the Medical Cost Personal Dataset (n=1,338), which includes demographic and lifestyle variables, this study examines the determinants of individual health insurance charges. A multivariate regression framework was used to estimate the effects of age, body mass index (BMI), smoking status, number of children, sex, and region. The results indicate that smoking increased the predicted charges by more than $23,000, while age and BMI also had statistically significant positive effects. In contrast, sex and regional differences were not significant after controlling for other factors. The model explained approximately 74% of the variance in charges. These findings align with prior evidence on lifestyle-related risks and provide support for risk-adjusted premium design, highlighting the policy relevance of smoking cessation and obesity prevention.

Keywords

health insurance charges, multivariate regression, smoking, body mass index, risk-based pricing, risk adjustment

1. Introduction

According to the Centers for Medicare & Medicaid Services (CMS), U.S. national health expenditures reached USD 4.5 trillion in 2022, accounting for 17.3% of gross domestic product (GDP), and are projected to grow at an average annual rate of 5.4% through 2031 [1]. Rising healthcare costs have placed sustained pressure on households, insurers, and policymakers. Health insurance plays a critical role in providing financial protection, yet premium growth and medical charges continue to outpace wage growth and inflation. Identifying the factors that drive variation in individual insurance charges is therefore essential for designing risk-adjusted premium structures and for informing effective public health strategies.

Prior research has consistently highlighted the influence of lifestyle and demographic characteristics on healthcare expenditures. Smoking has long been recognized as a major driver of costs. Evidence from national surveys shows that current smokers face substantially greater annual medical spending than nonsmokers do [2]. Moreover, longitudinal analyses confirm that smoking, along with other behavioral risks, leads to persistently elevated healthcare costs over time [3]. Obesity, which is commonly measured by body mass index

(BMI), is another significant determinant. Recent evidence from the Trøndelag Health Study in Norway demonstrated that higher BMI is consistently associated with greater healthcare expenditures, with costs rising sharply among individuals at the upper end of the BMI distribution [4]. Similar patterns are reported in U.S. data, where obesity has been estimated to account for approximately $147 billion in annual medical spending [5]. Age also exerts a fundamental influence, reflecting the growing burden of chronic conditions and long-term care needs among older populations. CMS data show that individuals aged 65 and older, who constitute 17% of the population, account for 37% of total personal healthcare spending [6], underscoring the fiscal significance of aging.

In addition to individual risk factors, regional and systemic variation also shape expenditure patterns. Comparative evidence from Germany has shown that regional differences in healthcare costs can be largely explained by variations in morbidity and the distribution of medical resources [7]. In the United States, the Dartmouth Atlas has documented large geographic variations in Medicare spending that are not consistently associated with better health outcomes, suggesting that contextual and provider-level factors drive cost disparities [8]. Moreover, population-level estimates indicate that modifiable behavioral risks such as smoking and high BMI account for a large share of total health spending, underscoring the policy importance of preventive strategies and actuarial risk adjustment [9].

Despite these insights, much of the evidence to date is derived from large-scale surveys or aggregate expenditure accounts, which may obscure variation at the individual level. Few studies apply a unified regression framework to jointly examine demographic, behavioral, and regional determinants via microlevel insurance data. This paper addresses that gap by analyzing the Medical Cost Personal Dataset (n = 1,338), a widely used dataset containing individual-level information on health insurance charges and related covariates. A multivariate regression model was applied to quantify the relative contributions of smoking, BMI, and age while also evaluating the additional roles of sex, number of children, and region. The study contributes by (1) confirming the effects of established cost drivers within a transparent regression framework, (2) assessing the significance of additional demographic and contextual variables, and (3) drawing implications for risk-adjusted premium design and public health interventions. In this way, it illustrates the value of accessible microdata for replicable empirical analysis and policy discussion in health economics.

2. Literature Review

2.1 Smoking and Healthcare Expenditures

Smoking is a major driver of excess medical costs because of its association with chronic diseases. U.S. survey data show that smokers face substantially higher annual expenditures than nonsmokers do [2]. National estimates attribute approximately 8–9% of total healthcare spending to smoking, exceeding USD 170 billion annually [10]. International evidence reinforces this burden, with the WHO reporting that smoking accounted for 5.7% of global health spending in 2012 [11]. Longitudinal analyses further indicate that reductions in smoking prevalence translate into measurable cost savings [3]. These findings underscore the economic importance of tobacco control and justify the inclusion of smoking status in expenditure models.

2.2 Body Mass Index and Obesity

Obesity, measured by BMI, is another key determinant of healthcare expenditures. U.S. estimates suggest that obesity adds over USD 140 billion in annual medical costs [5]. Using instrumental variables, Cawley and Meyerhoefer [12] showed that conventional estimates may understate the causal effect, with obesity increasing costs by approximately USD 2,700 per adult per year. More recent analyses confirm a national burden exceeding USD 170 billion annually, with severe obesity driving disproportionate costs [13]. OECD countries also report that overweight- and obesity-related conditions account for a sizable share of health budgets [14]. Collectively, these studies confirm that BMI is a critical predictor of both individual and system-level costs.

2.3 Age and Demographic Determinants

Age is a fundamental predictor, with per capita healthcare spending rising sharply among older adults. In the U.S., individuals aged 65+ account for 37% of total personal health spending, despite representing 17% of the population [6]. However, classic health economics research emphasizes that much of this association

reflects proximity to death rather than age per se [15]. Other demographic factors have mixed effects: sex-related differences often disappear once morbidity is controlled, whereas the number of dependent children tends to increase family expenditures but has a limited impact on per capita charges.

2.4 Regional Variation and Systemic Factors

Geographic variation also contributes to expenditure differences. In Germany, disparities are largely explained by morbidity and resource distribution [7]. In the U.S., Medicare spending varies by more than 40% across regions, yet higher spending does not consistently yield better outcomes [8, 16]. These findings suggest that systemic factors—provider practice styles, resource allocation, and postacute care utilization—are key drivers of regional variation, with implications for both insurers and policymakers.

2.5 Gaps in the Literature

Despite extensive evidence, three gaps remain. First, most analyses rely on aggregate survey or expenditure data, limiting insight into individual-level heterogeneity. Second, relatively few studies have jointly evaluated demographic, behavioral, and regional factors within a unified econometric framework. Third, while the costs of modifiable risks such as smoking and obesity are well established, their interactions with demographic variables and implications for premium design are less frequently addressed. This study contributes by applying multivariate regression to microlevel insurance data, offering a transparent assessment of established determinants and their relative importance.

3. Methodology

3.1 Data Sources

This study employs the Medical Cost Personal Dataset (N=1,338), which is publicly available and widely used in methodological demonstrations of health expenditure modeling. The dataset contains individual-level information on annual medical insurance charges and key demographic and lifestyle variables. While not a nationally representative survey, it provides a transparent microlevel setting that allows for the testing of established hypotheses regarding healthcare cost determinants.

3.2 Variables

The dependent variable is annual health insurance charges (USD). The independent variables include:

Age: measured in years.

Sex: binary indicator (male = 1, female = 0).

Body mass index (BMI): a continuous variable, defined as weight in kilograms divided by height in meters squared.

Children: number of dependent children covered by the policy.

Smoker: binary indicator (smoker = 1, nonsmoker = 0).

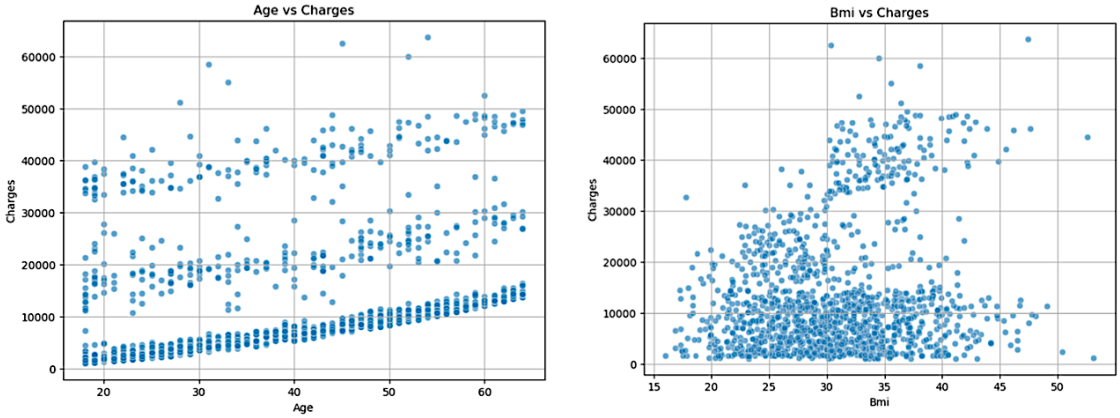
Region: categorical variable for U.S. geographic regions (Northeast, Northwest, Southeast, Southwest).

The selection of these variables is consistent with prior studies that identified demographic and lifestyle factors as key drivers of medical expenditures [2, 4, 17].

3.3 Descriptive Analysis

To provide an initial overview, scatter plots were generated to illustrate the bivariate relationships between charges and two major predictors: age and BMI (Figure 1). Charges increase steadily with age, whereas BMI also has a positive, although more dispersed, association. These patterns provide preliminary justification for including age and BMI as explanatory variables in the regression framework.

Figure 1: Scatter plots of annual health insurance charges by age and BMI



Scatter plot of age vs charges

Scatter plot of bmi vs charges

3.4 Statistical Considerations

Prior to estimation, descriptive statistics and graphical analyses were used to identify potential outliers and explore bivariate associations. Pairwise correlations were examined to assess multicollinearity among predictors, and residual plots were inspected to evaluate normality and heteroskedasticity. While the dataset does not permit advanced robustness checks such as instrumental variable estimation or longitudinal analysis, an ordinary least squares (OLS) framework provides a transparent baseline for assessing relative associations.

The regression model is specified as follows:

$$Charges_i = \beta_0 + \beta_1 Age_i + \beta_2 BMI_i + \beta_3 Children_i + \beta_4 Smoker_i + \beta_5 Sex_i + \beta_6 Region_i + \epsilon_i$$

where $Charges_i$ represents annual medical charges for individual i and where ϵ_i is the error term.

4. Results

4.1 Descriptive Statistics by Region

Table 1 reports summary statistics of annual insurance charges across four U.S. regions. The Southeast region has the highest mean charges (USD 63,770) and moderate variability, indicating both higher overall costs and dispersion. Southwest China has the lowest mean (USD 52,591) but the highest standard deviation, suggesting considerable heterogeneity in expenditures within the region. In comparison, Northeast China and Northwest China display more consistent patterns, with moderately high means and relatively lower standard deviations. These regional differences may reflect variations in population health, provider availability, or local economic conditions.

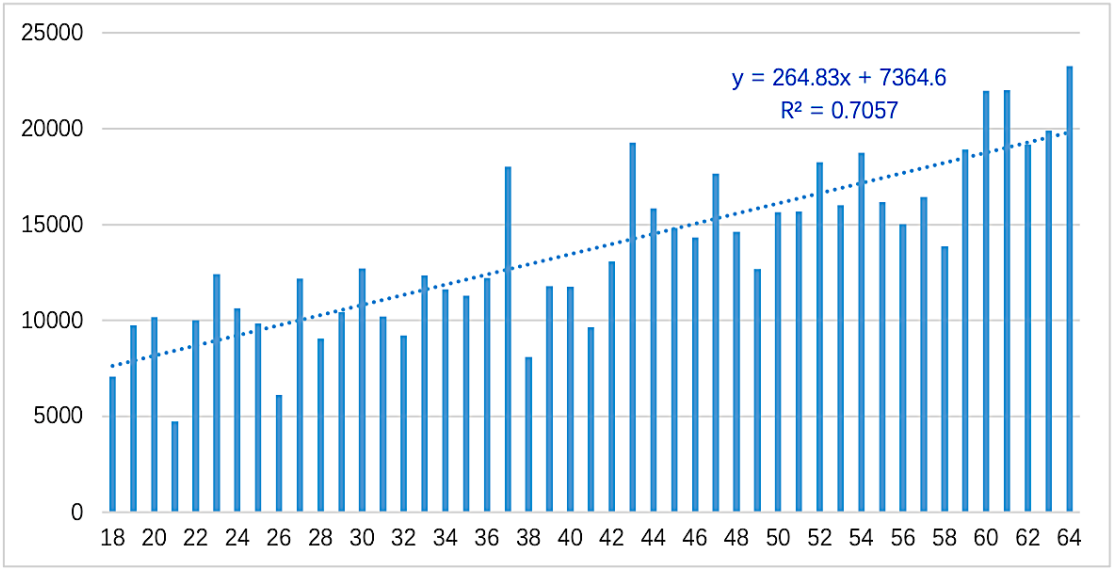
Table 1. Descriptive statistics of annual insurance charges by region

Region	Number	Sum	Mean	Standard deviation
Northeast	324	4343669	58571	11256
Northwest	325	4035712	60021	11072
Southwest	325	4012755	52591	13971
Southeast	364	5363690	63770	11557

4.2 Age and Healthcare Charges

Figure 3 depicts average annual charges by age, with a fitted linear trend line. A positive association is observed, with charges rising by approximately USD 265 per year of age ($R^2 = 0.71$). While the overall trend is increasing, fluctuations are evident, particularly between the ages of 30 and 50, suggesting heterogeneity attributable to unobserved health or behavioral factors.

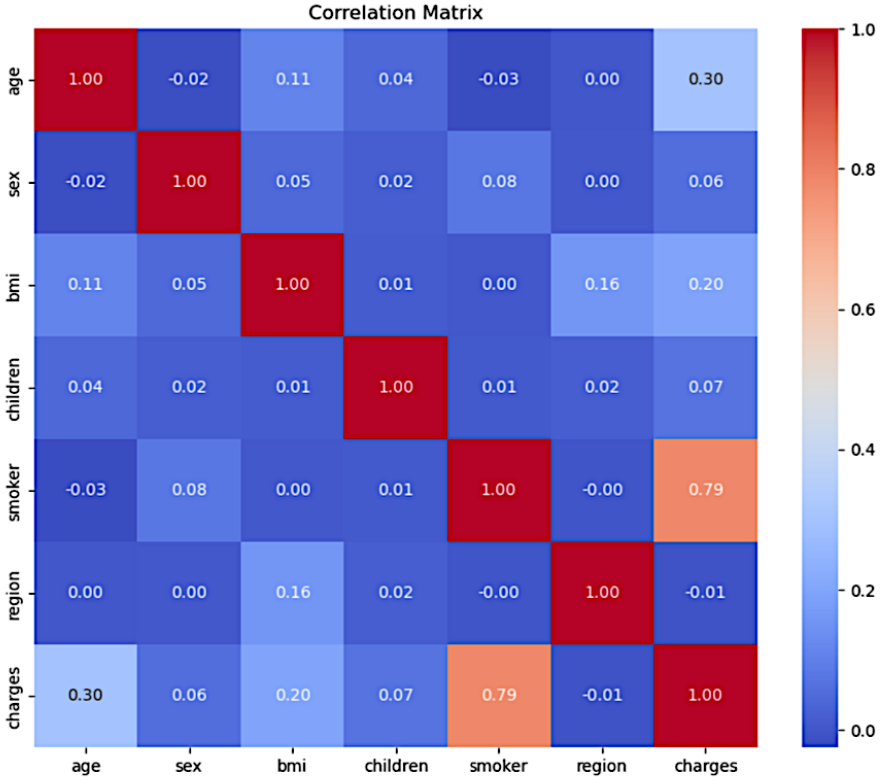
Figure 2: Average annual insurance charges by age with a fitted trend line



4.3 Correlation Analysis

Figure 4 presents the correlation matrix of the key variables. Smoking status had the strongest correlation with charges (r = 0.79), followed by age (r = 0.30) and BMI (r = 0.20). Other variables, such as sex and region, exhibited negligible associations. These patterns provide preliminary evidence that lifestyle-related factors dominate demographic and geographic determinants in shaping insurance costs.

Figure 3: Correlation matrix of demographic, lifestyle, and regional variables



4.4 Regression Analysis

Table 2 reports the F test results for the regression model. The overall F statistic of 508.7 ($p < 0.001$) confirms that the independent variables jointly explain a significant share of the variation in charges. The model achieves an R^2 of 0.742 (adjusted R^2= 0.740), indicating that approximately 74% of the variance in annual insurance charges is explained by the predictors.

Table 2: F test results for the regression model

Source	DF	FValue	P-Value
Model	6	508.7	<.0001
Error	1063		
Total	1069		

Table 3 presents the estimated coefficients. Smoking status has the largest effect, with smokers incurring USD 23,650 higher annual charges than nonsmokers do ($p < 0.001$). Age and BMI also have significant positive effects, with each additional year of age associated with USD 257 higher charges and each unit increase in BMI linked to USD 336 higher charges. The number of children is positively associated with charges (USD 425 per child, $p < 0.01$). In contrast, sex and regional indicators are statistically insignificant, suggesting a limited influence on individual-level variation once behavioral risk is controlled for.

Table 3. Parameter estimates from the multivariate regression model

Variable	Parameter Estimate	Standard Error	T Value	P Value
Intercept	-1.195e+04	1086.94	-10.991	<.0001
Age	257.06	13.45	19.109	<.0001
Sex	-18.79	375.77	-0.05	0.96
BMI	335.78	31.66	10.607	<.0001
Children	425.09	154.43	2.753	0.006
Smoker	2.365e+04	465.25	50.829	<.0001
Region	-271.28	170.37	-1.592	0.112

Figure 4 compares the predicted and actual charges. Most observations lie close to the 45-degree line, confirming good model fit, although larger deviations are visible at higher charge levels. Figure 5 shows the residual distribution, which approximates normality but reveals heavier tails driven by smokers and individuals with elevated BMI.

Figure 4: Actual versus predicted charges from the regression model

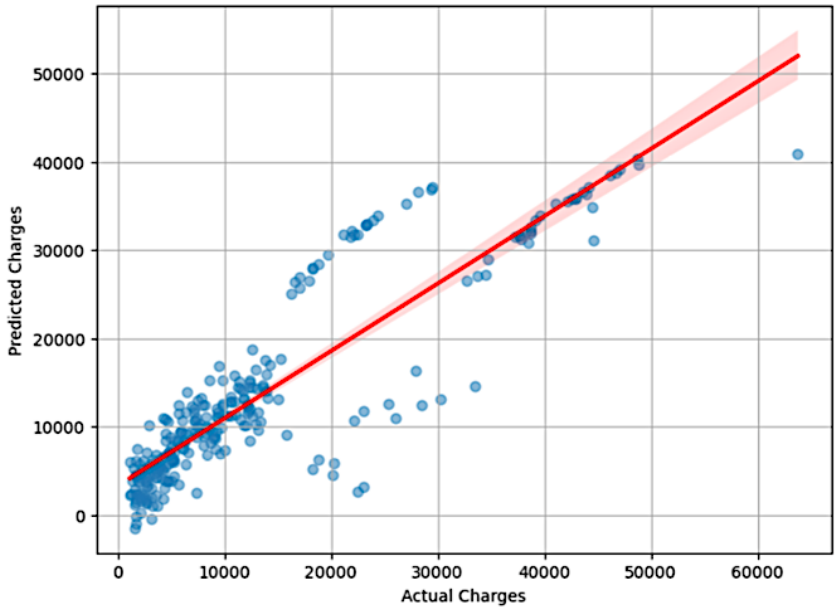
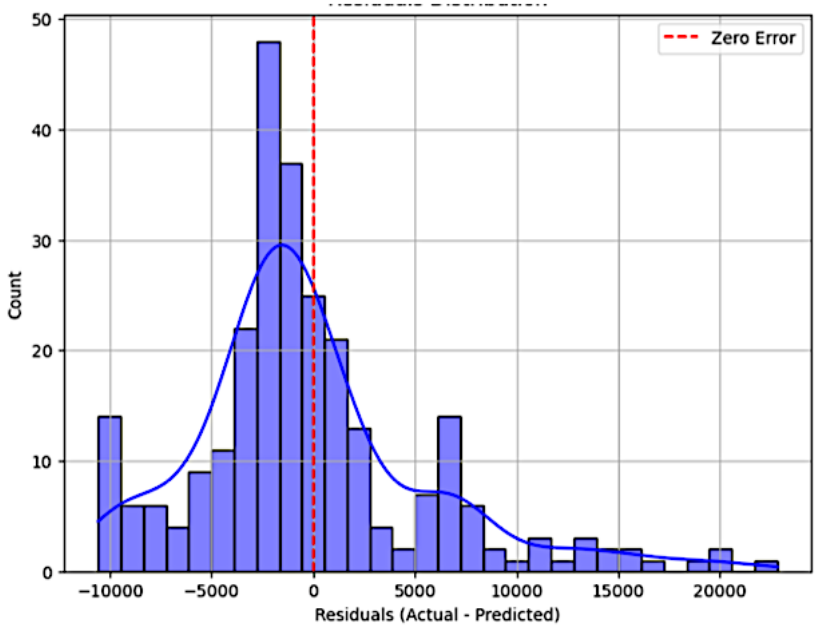


Figure 5: Residual distribution of the regression model



5. Conclusion

This study examined the determinants of U.S. health insurance charges via microlevel data and a multivariate regression framework. The results demonstrate that smoking is by far the most influential predictor of charges, followed by age and BMI, whereas sex and regional indicators are not statistically significant. The model explains approximately 74% of the variance in charges, underscoring the importance of behavioral risk factors in shaping medical expenditures. These findings are consistent with recent evidence showing that smoking continues to impose a substantial burden on U.S. healthcare spending [18, 19] and that obesity-related multimorbidity significantly increases costs across demographic groups [20]. Longitudinal analyses further confirm that lifestyle risks such as smoking and obesity independently drive higher medical expenditures [17], whereas demographic and regional differences explain relatively little variation once risk factors are controlled for [21].

The findings have direct implications for policy and practice. For insurers, incorporating behavioral risk factors such as smoking and BMI into premium design can improve pricing accuracy. For policymakers, targeted interventions aimed at smoking cessation and obesity prevention could substantially reduce overall healthcare costs and improve public health outcomes.

Several limitations should be noted. The dataset, while useful for methodological demonstration, is not nationally representative and lacks detailed clinical information. The cross-sectional design also precludes analysis of dynamic changes over time. Future research should incorporate longitudinal and nationally representative data and explore additional explanatory factors, such as genetic predispositions, socioeconomic status, and detailed behavioral measures.

Overall, the analysis highlights that modifiable lifestyle risks are central drivers of health insurance charges. By prioritizing prevention and integrating risk adjustment mechanisms, both insurers and policymakers can move toward a more equitable and sustainable approach to managing healthcare costs.

References

[1] Centers for Medicare & Medicaid Services (CMS). National health expenditure projections 2022–2031. Baltimore, MD: U.S. Department of Health and Human Services, 2023.

[2] Swedler, D. I., Miller, T. R., Ali, B., Waeher, G. and Bernstein, S. L. National medical expenditures by smoking status in American adults: an application of Manning's two-stage model to nationally representative data. BMJ open. 2019, 9(7), p. e026592. https://doi.org/10.1136/bmjopen-2018-026592.

[3] Lightwood, J. and Glantz, S. A. Smoking behavior and healthcare expenditure in the United States, 1992–2009: panel data estimates. PLoS medicine. 2016, 13(5), p. e1002020. https://doi.org/10.1371/journal.pmed.1002020.

[4] Hansen Edwards, C., Håkon Bjørngaard, J., Minet Kinge, J., Åberge Vie, G., Halsteinli, V., Ødegård, R., Kulseng, B. and Waaler Bjørnelv, G. The healthcare costs of increased body mass index-evidence from The Trøndelag Health Study. Health Economics Review. 2024, 14(1), p. 36. https://doi.org/10.1186/s13561-024-00512-8.

[5] Finkelstein, E. A., Trogdon, J. G., Cohen, J. W. and Dietz, W. Annual Medical Spending Attributable To Obesity: Payer-And Service-Specific Estimates: Amid calls for health reform, real cost savings are more likely to be achieved through reducing obesity and related risk factors. Health affairs. 2009, 28(Suppl1), pp. w822-w831. https://doi.org/10.1377/hlthaff.28.5.w822.

[6] Centers for Medicare & Medicaid Services (CMS). National health expenditure data: Personal health care spending by age and gender, 2010–2020. Baltimore, MD: U.S. Department of Health and Human Services, 2020.

[7] Göpffarth, D., Kopetsch, T. and Schmitz, H. Determinants of regional variation in health expenditures in Germany. Health economics. 2016, 25(7), pp. 801-815. https://doi.org/10.1002/hec.3200.

[8] Skinner, J. and Fisher, E. S. Reflections on geographic variations in U.S. health care. Dartmouth Atlas White Paper. Hanover, NH: Dartmouth Institute for Health Policy & Clinical Practice, 2010.

[9] Bolnick, H. J., Hodgson, T. A. and Caplan, C. Health and economic impact of chronic diseases and modifiable risk factors in the United States. Journal of Public Health Management and Practice. 2020, 26(2), pp. 92-99. https://doi.org/10.1097/PHH.0000000000001100.

[10] Xu, X., Bishop, E. E., Kennedy, S. M., Simpson, S. A. and Pechacek, T. F. Annual healthcare spending attributable to cigarette smoking: an update. American journal of preventive medicine. 2015, 48(3), pp. 326-333. https://doi.org/10.1016/j.amepre.2014.10.012.

[11] Goodchild, M., Nargis, N. and d'Espaignet, E. T. Global economic cost of smoking-attributable diseases. Tobacco control. 2018, 27(1), pp. 58-64. https://doi.org/10.1136/tobaccocontrol-2016-053305.

[12] Cawley, J. and Meyerhoefer, C. The medical care costs of obesity: an instrumental variables approach. Journal of health economics. 2012, 31(1), pp. 219-230. https://doi.org/10.1016/j.jhealeco.2011.10.003.

[13] Ward, Z. J., Bleich, S. N., Long, M. W. and Gortmaker, S. L. Association of body mass index with health care expenditures in the United States by age and sex. PloS one. 2021, 16(3), p. e0247307. https://doi.org/10.1371/journal.pone.0247307.

[14] World Health Organization (WHO) and Organisation for Economic Co-operation and Development (OECD). The heavy burden of obesity: The economics of prevention. Paris: OECD Publishing. , 2019.

[15] Zweifel, P., Felder, S. and Meiers, M. Ageing of population and health care expenditure: a red herring? Health economics. 1999, 8(6), pp. 485-496. https://doi.org/10.1002/(SICI)1099-1050(199909)8:6<485::AID-HEC461>3.0.CO;2-4.

[16] Institute of Medicine. Variation in Health Care Spending: Target Decision Making, Not Geography. Washington, DC: National Academies Press, 2013.

[17] Kim, Y. The effects of smoking, alcohol consumption, obesity, and physical inactivity on healthcare costs: a longitudinal cohort study. BMC Public Health. 2025, 25(1), p. 873. https://doi.org/10.1186/s12889-025-xxxx.

[18] Gu, D., Sung, H.-Y., Calfee, C. S., Wang, Y., Yao, T. and Max, W. Smoking-attributable health care expenditures for US adults with chronic lower respiratory disease. JAMA Network Open. 2024, 7(5), pp. e2413869-e2413869. https://doi.org/10.1001/jamanetworkopen.2819203.

[19] Nargis, N., Hussain, A. G., Asare, S., Xue, Z., Majmundar, A., Bandi, P., Islami, F., Yabroff, K. R. and Jemal, A. Economic loss attributable to cigarette smoking in the USA: an economic modelling study. The Lancet Public Health. 2022, 7(10), pp. e834-e843. https://doi.org/10.1016/S2468-2667(22)00020-X.

[20] Ezendu, K., Pohl, G., Lee, C. J., Wang, H., Li, X. and Dunn, J. P. Prevalence of obesity-related multimorbidity and its health care costs among adults in the United States. Journal of Managed Care & Specialty Pharmacy. 2025, 31(2), pp. 179-188. https://doi.org/10.18553/jmcp.2025.31.2.179.

[21] Adjei, N. N., Haas, A., Sun, C. C., Zhao, H., Yeh, P. G., Giordano, S. H., Toumazis, I. and Meyer, L. A. Healthcare Costs in the United States by Demographic Characteristics and Comorbidity Status. Value in Health. 2025, 28(2), pp. 206-214. https://doi.org/https://doi.org/10.1016/j.jval.2024.10.3847.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access

This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.