A Segment of Healthy and Unhealthy Lifestyle Consumers Affects Healthcare Expenditures: An Application of Data Mining in Healthcare

Hsin-Yen Yen, Ching Li, and Ping-Feng Hsia

***Abstract*—Data mining is a useful tool to analyze healthcare information and identify personal behavior patterns for policy decision-making., The purpose of this study is to identify healthy and unhealthy families using the household income and expenditure database and to analyze the difference of household healthcare expenditures among the healthy and unhealthy lifestyle groups. The database was composed of the data of the DGBAS surveys over 40 years. The methods in this study were descriptive analyses, ANOVA, cluster analysis, and the Chi-square test. The descriptive results showed there were four types of groups, including Smoker, Alcoholic, Unhealthy diet, and Healthy lifestyle. The healthcare expenditure of the Healthy lifestyle group was significantly lower than that of the other three unhealthy lifestyle groups. Keeping a healthy lifestyle is important for a family since it may decrease healthcare expenditure. The government has to be aware and reduce the cause of high healthcare expenditure by policy-making, to promote well-being, and create a better society.**

***Index Terms*—Cluster analysis, database, health promotion, household income, healthcare economic.**

I. Introduction

Data mining is a process of discovering knowledge, identifying new patterns and trends in databases, and both forecasting and predicting future events [1], [2]. Data mining has been used extensively, not only for information technology, but also for management, finances, social science, and healthcare. In the healthcare industry, data mining is used as a tool to analyze medical insurance claims, financial and clinical data, and health outcomes, to assist health-related decision-making as well as improve quality, efficiency, and efficacy. Overall, data mining is an approach to treatment effectiveness, healthcare management, consumer relationship management, and both fraud and abuse detection [3].

Data mining is the discipline generated by combining aspects of machine learning, artificial intelligence, statistics, and probability [4]. The most common approaches are descriptive analysis, predictive analysis, prescriptive analytics, and modeling, including classification, regression, association rule, cluster analysis, text mining, decision tree analysis, link analysis, and both detection n and identification algorithms [2], [5]. Therefore, this study investigated whether one of data mining skills, cluster analysis, could be applied to the healthcare research field by presenting meaningful value.

The household expenditure for healthcare has an impact on the development and economy of the country. It is also an important indicator of both health policies and strategies to ensure the equity, efficiency, quality, and cost of public healthcare services [6]. According to the Survey of Family Income and Expenditure in Taiwan, the proportion of healthcare expenditure has increased from 9.8% in 1996 to 15% in 2015, because of both prolonged life expectancy and additional health promotion [7]. On the other hand, if healthcare expenditure is catastrophic and unaffordable, it could thrust the household into poverty and seriously have an impact on the family's well-being [8]. Therefore, it is important to maintain a low proportion of healthcare expense as part of the overall household expenditure.

The healthcare expenditure relates to healthcare costs and out-of-pocket expenses. Healthcare expenditure includes payments to hospitals, physicians, dentists, and healthcare institutions, as well as fees for inpatients and outpatients (e.g., transportation, accommodations, and other medical instruments) [9]. Healthcare expenditure is affected by the GDP, the health insurance system, household income, and the over 65 years old population [10]. Low-income countries that had poor health insurance and health policy agendas presented a higher household healthcare expenditure [8].

From a disease prevention perspective, healthcare expenditure related to aging, individuals' unhealthy lifestyle, life expense, and diseases. The determinants of healthcare expenditure also correlated to injury and several non-communicable diseases of family members, such as diabetes, asthma, and heart disease [11]. The common risk factors of non-communicable diseases are the environment, demography, individual's behaviors, and unhealthy lifestyle [12]. Fig. 1 shows personal consumer behavior and other risk factors for disease [13].

Personal consumer behavior is an important determinant of disease. Lifestyle is a self-determining process. Cognitive health status, the definition of health, and self-concept are elements of a health-promoting lifestyle [14]. Healthy lifestyles emphasize that an individuals' behavior is to minimize health problems and to maximize their own well-being. Individuals who engage in a healthy lifestyle seek a healthy diet and change their consumer behaviors by increasing their intake of vegetables and fruits, low-fat foods, low carbohydrate drinks, and organic foods, while avoiding unhealthy products, like alcohol, soft drinks, and processed

Manuscript received December 12, 2016; revised April 12, 2017.

Hsin-Yen Yen, Ching Li, and Ping-Feng Hsia are with Graduate Institute of Sport, Leisure, and Hospitality Management, National Taiwan Normal University, 106, Taipei City, Taiwan (e-mail: kenjizoro520@gmail.com, t94002@ntnu.edu.tw, 60031011a@ntnu.edu.tw).

doi: 10.18178/ijfcc.2017.6.3.495

snacks [15]. An individual's eating patterns, nutrient intake, and food preferences also affect the onset of metabolic diseases [16]. In addition, the control of tobacco and alcohol use decreases the risk of non-communicable diseases [11].

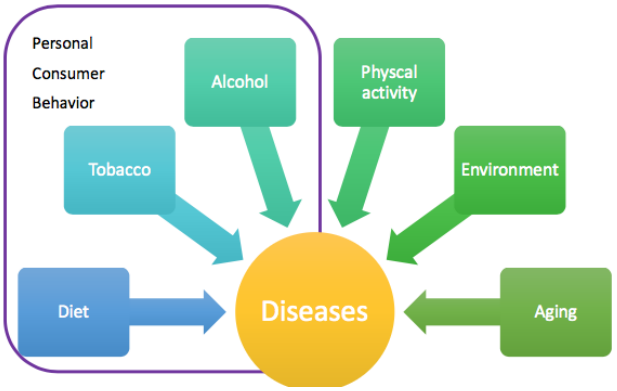

Fig. 1. Risk factors for disease.

The lifestyle of both the individual and the family affect healthcare expenditure. People who have a healthy-living lifestyle, have a lower healthcare expenditure [17], [18]. Alcohol and cigarettes have a positive association with household healthcare expenditure [19]. In order to maintain a healthy lifestyle, individuals modify their purchasing behavior, including food choices, eating patterns, and alcohol or cigarettes use.

Previous studies about health, food, alcohol, or tobacco use of household expenditure were usually analyzed and discussed independently of demographic characteristics [20]–[23]. Therefore, the purposes of this study to identify the healthy and unhealthy families through the database, to understand the demographic characteristics of healthy and unhealthy families and their unhealthy consumer behavior, and to analyze the difference of household healthcare expenditure among the healthy and unhealthy lifestyle groups.

II. Method

A. Database

"Report on the Survey of Personal Incomes" had started in 1964. Since 1994, the survey has been revised, and the latest version is the "Report on the Survey of Family Income and Expenditure Distribution in Taiwan Area," which was investigated by the Directorate-General of Budget, Accounting, and Statistics (DGBAS), Executive Yuan, Taiwan. The purpose of this series of surveys is to understand the household income and expenditure in the Taiwan area and to estimate the price index, distribution, savings, and consumer patterns. The government and organizations could use this report as a reference and refer to it as needed to develop a social plan, improve citizens' lives, and promote social welfare [7].

Every year, a two-stage sampling method with counties and cities as subpopulations is applied. The sampling rate is about 0.20%, which are 16,528 households in 2015. Households are required to do account keeping for a period of one year. The process of data collection is done by interviewing households from January to the following February. The items of the survey are shown in Fig. 2 and include household members, household facilities and housing conditions, income and expenditure, and consumption expenditure.

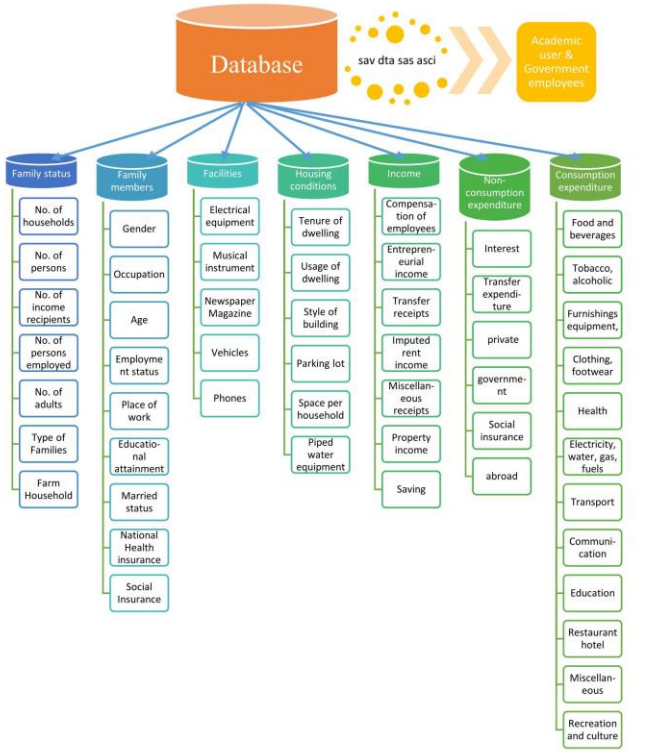

Fig. 2. Content of the databases.

All the results of the surveys from 1977 to 2016 are integrated into a database. The database was constructed by Survey Research Data Archive (SRDA), Center for Survey Research, Academia Sinica for an academic purpose. The database renews once in October every year to release the data of the previous year. The database supplies survey-related documents, such as the instructions, the questionnaire, coding book, and report. Furthermore, the format of the original data supplies in sav, dta, sas, and asci. Access to the database is free for registered academic users, including students, teachers, professors, faculty of academic institutions, and employees of Academia Sinica or government departments.

B. Data Analysis

In this study, the data of household members, income, and consumption expenditure in the database were analyzed. Descriptive analysis, ANOVA, cluster analysis, and a Chi-square test were performed to analyze the data using SPSS software. The items of consumption expenditure used were:

1) Glucose products: sugar, candy, jam, honey, chocolate, ice cream, etc.
2) Non-alcoholic beverages: soda, root beer, coke, juice, tea, cocoa, etc.
3) Eating out: restaurants, food vendors, cafeteria, take-out food, etc.
4) Tobacco products and betel nuts
5) Alcoholic beverages: beer, wine, cocktails, sorghum liquor, rice wine, etc.

6) Medical and healthcare products: medical devices and equipment (e.g., glasses, manometer, contacts, prosthesis), medical care (e.g., inpatients, outpatients, dental, alternative medicine), pharmacy (e.g., OTC-drug, vitamin, Chinese pharmacy), and National Health Insurance.

III. Result

A. Cluster Analysis

In 2015, a total of 16527 households had data on household incomes and expenditures. All households in the database were divided into four patterns of consumers. Fig. 3 illustrates the four patterns and their corresponding numbers, including (1) Alcoholic, (2) Smoker, (3) Unhealthy diet, and (4) Healthy lifestyle.

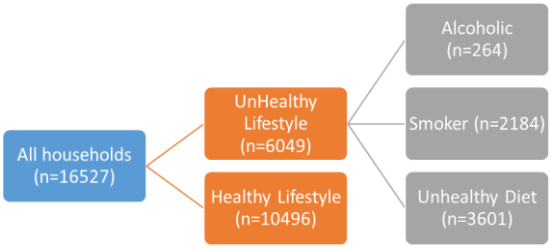

Fig. 3. Results of cluster analysis.

Table I summarizes the results of the cluster analysis. There were significant differences of the Z-score of all unhealthy items among four patterns of consumers (p < 0.000***), including glucose products, non-alcoholic beverages, eating out, tobacco products and betel nuts, and alcoholic drinks. The family in the first group, Alcoholic (n = 246), preferred alcohol drinks and tobacco products also spent significantly more for alcoholic drinks than did the other households. The family in the second group, Smoker (n = 2184), spent more for tobacco and betel nuts than other households. The family in the third group, Unhealthy Diet (n = 3601), preferred unhealthy food and eating out. The household expenditure of glucose products, non-alcoholic beverages, and eating out in the Unhealthy Diet groups was significantly higher than other groups. Finally, the family in the fourth group presented a healthy lifestyle (n = 10496) and had the lowest expenditures of unhealthy food, bad eating behavior, tobacco products, and alcoholic beverages.

B. The Characteristics of Healthy and Unhealthy Families

Table II summarizes the characteristics of households in each group. Overall, the average number of family members is 3.089, the average number of minor family members is 0.569, and the average number of elder family members is 0.589. The distribution of family members in Alcoholic, Smoker, and Unhealthy diet groups is similar among all households. However, the number of minor family members is lower, and the number of elder family members is higher than other groups. Furthermore, household income ranged between 19,077 and 23,602,409 new Taiwan dollars and averaged 1,163,162 in 2015.

TABLE I: Results of Cluster Analysis

	Cluster	glucose products	non-alcoholic	eating out	tobacco	alcoholic drink
1	Alcoholic	0.25	0.71	0.67	1.67	5.22
2	Smoker	−0.11	0.06	0.23	1.89	0.15
3	Unhealthy diet	1.12	0.99	0.87	−0.19	0.08
4	Healthy lifestyle	−0.37	−0.37	−0.36	−0.37	−0.19
	F	3099.24	2432.79	1945.30	8500.22	7525.06
	p	0.00***	0.00***	0.00***	0.00***	0.00***

TABLE II: Descriptive Characteristics of the Four Groups

Cluster		No. of Family members	No. of under age	No. of elder family	Household income	Age of household head
1 Alcoholic	Mean	3.70	0.76	0.33	1853743	50.37
	SD	1.95	1.04	0.62	1979987	10.51
2 Smoker	Mean	3.62	0.71	0.46	1221129	47.95
	SD	1.55	0.95	0.71	652569	11.57
3 Unhealthy diet	Mean	4.16	0.97	0.46	1764706	47.39
	SD	1.36	1.04	0.72	989105	10.97
4 Healthy lifestyle	Mean	2.60	0.40	0.67	928536	54.90
	SD	1.25	0.75	0.79	629114	15.97
Total	Mean	3.09	0.57	0.59	1163162	52.28
	SD	1.49	0.88	0.77	836774	14.82

The characteristics of household heads in each group are shown in Table III. The gender of household heads was male more than female in each group and total households [χ^2 (4) = 599.934, p < 0.000***]. The average age of household heads was 52.48 years. The age of household heads in Alcoholic and Healthy lifestyle groups was higher than that of both Smokers and Unhealthy diet groups. [F (3, 16523) = 321.455, p < 0.000***] As for the education level of household heads, the frequency level in the Alcoholic and Smoker groups was below high school. The frequency education level in the Unhealthy diet group was between high school and college,

and this group had the highest frequency at the graduate level. The frequency education level in the Healthy Lifestyle group was between none and college [$\chi^2(4) = 1314.409$, $p < 0.000$***].

The frequency of family structures was the orderly nuclear family (35.20%), the couple (19.04%), three generations (14.21%), one-person (12.14%), and single parent (10.30%) in all households. The frequency of family structures in Unhealthy lifestyle groups was similar to those of all households. However, nuclear family (26.74%) and the couple (25.35%) were both the main types of family structures in the Healthy Lifestyle groups [$\chi^2(4) = 2899.533$, $p < 0.000$***].

The occupations of household heads were ordered none, manufacture, sales and retail, construction, agriculture, and hospitality. The more frequent occupations of household heads in the Alcoholic and Smoker groups were manufacture and construction. The more frequent occupations of household heads in the Unhealthy diet groups were manufacture and construction. Nevertheless, the most frequent occupation of household heads in the Healthy lifestyle group was no job [$\chi^2(4) = 2172.114$, $p < 0.000$***].

TABLE III: CHARACTERISTICS OF HOUSEHOLD HEADS OF THE FOUR GROUPS

Cluster	1		2		3		4		Total	
	N	%	N	%	N	%	N	%	N	%
Gender										
male	212	86.18	1857	85.03	2827	78.51	6740	64.22	11636	70.41
female	34	13.82	327	14.97	774	21.49	3756	35.79	4891	29.59
Education Level										
none or elementary	90	36.59	809	37.04	603	16.75	3934	37.48	5436	32.89
high school	86	34.96	863	39.52	1132	31.44	2945	28.06	5026	30.41
college	55	22.36	467	21.38	1457	40.46	3075	29.30	5054	30.58
graduate	15	6.10	45	2.06	409	11.36	542	5.16	1011	6.12
Occupation										
none	12	4.88	140	6.41	146	4.05	3136	29.88	3434	20.78
manufacture	65	26.42	518	23.72	946	26.27	1849	26.42	518	23.72
sale and retails	35	14.23	263	12.04	524	14.55	1117	14.23	263	12.04
construct	44	17.89	394	18.04	357	9.91	613	5.84	1408	8.52
agriculture	11	4.47	129	5.91	84	2.33	610	5.81	834	5.05
hospitality	12	4.88	137	6.27	150	4.17	470	4.48	769	4.65
services	11	4.47	87	3.98	168	4.67	407	3.88	673	4.07
transportation	10	4.07	139	6.36	158	4.39	321	3.06	628	3.80
public services	13	5.29	78	3.57	193	5.36	327	3.12	611	3.70
education	5	2.03	29	1.33	167	4.64	354	3.37	555	3.36
Family Structure										
one-person	22	8.94	119	5.45	12	0.33	1854	17.66	2007	12.14
couple	29	11.79	247	11.31	209	5.80	2661	25.35	3146	19.04
single Parent	21	8.54	228	10.44	192	5.33	1262	12.02	1703	10.30
nuclear family	96	39.02	899	41.16	2015	55.96	2807	26.74	5817	35.20
grandparent-child	1	0.41	18	0.82	12	0.33	169	1.61	200	1.21
three generations	53	21.55	479	21.93	921	25.58	895	8.53	2348	14.21
others	24	9.76	194	8.88	240	6.67	848	8.08	1306	7.90

TABLE IV: RESULTS OF THE ANOVA

	Sum of Squares	*df*	Mean Square
Between Groups	7491852424159.83	3	2497284141386.61
Within Groups	200780379388853.00	16523	12151569290.62
Total	425577400581800.00	16527	
F	205.51		
p	0.000***		
post-hoc	1,3>2>4		

C. The Difference of Healthcare Expenditure

The healthcare expenditure of 1 to 4 groups was orderly 150893.48, 118062.88, 151890.50, and 100340.40. Table IV summarizes the results of the ANOVA among the four groups. There was a significant difference in household healthcare expenditure among the four groups [$F(3, 16523) = 205.511$, $p < 0.000$***]. Moreover, the household healthcare expenditure in the Alcoholic and Unhealthy diet groups was significantly higher than the Smoker group, and all unhealthy lifestyle consumers' household healthcare expenditures were significantly higher than were those of healthy lifestyle consumers.

IV. DISCUSSION

Using the evidence provided by the household income and expenditure database enabled the identification of healthy and unhealthy families based on the consumer behavior of households. The consumer behavior of an unhealthy family presented an unbalanced diet, and use of both alcohol and tobacco products. On the other hand, the healthy lifestyle family spent less money on sugary products, alcoholic beverages, and tobacco products than did the unhealthy family. The previous study also divides the families into healthy and unhealthy groups, according to the expenditure of alcoholic beverages, food, soft drinks, and glucose

products [15].

The alcoholic family was mainly a nuclear family and had the highest household income and the fewest number of elderly members. The smoker family had both a similar family situation and family structure to the alcoholic family but had a lower household income. The unhealthy diet family was also mainly a nuclear family and had the highest number of family members and minors. The family structure of the healthy lifestyle family was equivalent to a nuclear family, which had the least number of family members, but the highest number of elderly and the oldest age of household heads. Although aging has an impact on both personal and household expenditure, elder family members affect consumer behavior to become a healthy lifestyle [17].

This study determined there was a difference in healthcare expenditure between healthy and unhealthy families. More expenditure on sugary products, beverages, eating out behavior, tobacco products, and alcoholic beverages could lead to increased healthcare costs. An unhealthy lifestyle might threaten our health status that ultimately forces us to become an economic burden on our household [12]. The money spent on alcohol or cigarettes has a positive association with household healthcare expenditure [19].

Despite this, there were still several limitations in this study. First, in the process, before data enter into the database, part of data collection method required participants to recall expenditure from their memory. In this case, the bias of both real data and missing data could exist. Second, the healthcare expenditure for the type of diseases is unknown. The hypothesized variable was regular healthcare expenditure. The healthcare cost of communicable diseases and accidents should be excluded. Third, the healthy lifestyle omitted a factor, physical activity, but this behavior should be considered. Moreover, the number of family members who engage in unhealthy behavior is also unknown. For example, tobacco expenditure does not represent all family members who may smoke [20]. In future studies, these limitations could be overcome by addressing them beforehand. The database could be analyzed by individuals' behavior, according to both personal income and expenditure. Other healthy lifestyle items could be added into variables as well, such as sport expense, recreational expense, and social activity. Moreover, the survey in this database has been over 20 years. The longitudinal study design could be used for the change of consumer behavior.

V. Conclusion

This study identified a segment of the healthy and unhealthy lifestyle family groups that caused a different result in healthcare expenditure. An unhealthy lifestyle increases not only the incidence of disease but also the economic burden of the household [12]. A healthy lifestyle should be engaged in by their own. For instance, before purchasing food and drink, ingredients and nutrition facts should be considered, especially glucose content. Furthermore, both policy makers and health systems should pay close attention to the issue of rising healthcare expenditure [8]. The regulation of tobacco and alcohol has to be enhanced coupled with both increased prices and taxes to reduce tobacco use and alcohol consumption [11]. Healthy lifestyle is important to promote well-beings. Furthermore, this study used data mining applications to recognize health issues. Data mining is a powerful tool and could be applied in social science research as well.

References

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, pp. 37–54, 1996.

[2] Ț. Ionut, "Data mining in healthcare: Decision making and precision," *Database Syst. J.*, vol. VI, no. 4, pp. 33–40, 2015.

[3] H. C. Koh and G. Tan, "Data mining applications in healthcare," *J Heal. Inf Manag*, vol. 19, no. 2, pp. 64–72, 2005.

[4] N. Jothi, N. A. Rashid, and W. Husain, "Data Mining in Healthcare — A review," *Procedia Comput. Sci.*, vol. 72, pp. 306–313, 2015.

[5] B. D. Crockett, R. Johnson, and B. Eliason, "What is data mining in healthcare ?" pp. 1–13, 2014.

[6] M. S. Haque and S. D. Barman, "Determinants of household healthcare expenditure in Chittagong, Bangladesh," *IUP J. Appl. Econ.*, vol. 9, no. 2, pp. 5–13, 2010.

[7] E. Y. Directorate, "The survey of family income and expenditure, 2015 (AA170040)," *Survey Research Data Archive, Academia Sinica*, 2016.

[8] K. Xu, D. B. Evans, K. Kawabata, R. Zeramdini, J. Klavus, and C. J. Murray, "Household catastrophic health expenditure: A multicounty analysis," *Lancet*, vol. 362, pp. 111–117, 2003.

[9] S. Mukherjee, S. Haddad, and D. Narayana, "Social class related inequalities in household health expenditure and economic burden: Evidence from Kerala, south India.," *Int. J. Equity Health*, vol. 10, no. 1, p. 1, 2011.

[10] U.-G. Gerdtham and B. Jönsson, "International comparisons of health expenditure: Theory, data and econometric analysis," *Handb. Heal. Econ.*, vol. 1, pp. 11–53, 2000.

[11] E. Saito, S. Gilmour, M. M. Rahman, G. S. Gautam, P. K. Shrestha, and K. Shibuya, "Catastrophic household expenditure on health in Nepal: A cross-sectional survey," *Bull. World Health Organ.*, vol. 92, no. 10, pp. 760–767, 2014.

[12] D. J. Hunter and K. S. Reddy, "Noncommunicable diseases," *N. Engl. J. Med.*, vol. 369, no. 14, pp. 1336–1343, 2013.

[13] World Health Organization. (2015). Non communicable diseases. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs355/en/

[14] D. E. King, A. G. Mainous, M. Carnemolla, and C. J. Everett, "Adherence to healthy lifestyle habits in us adults, 1988-2006," *Am. J. Med.*, vol. 122, no. 6, pp. 528–534, 2009.

[15] U. R. Orth, M. Mcdaniel, T. Shellhammer, K. Lopetcharat, R. L. Divine, and L. Lepisto, "Analysis of the healthy lifestyle consumer," *J. Consum. Mark. Br. Food J. Br. Food J. Iss J. Consum. Mark.*, vol. 22, no. 12, pp. 275–283, 2005.

[16] K. K. Li *et al.*, "An examination of sex differences in relation to the eating habits and nutrient intakes of university students," *J. Nutr. Educ. Behav.*, vol. 44, no. 3, pp. 246–250, 2012.

[17] B. Dormont, M. Grignon, B. Dormont, and M. Grignon, "Health expenditure growth : Reassessing the threat of ageing To cite this version : Health expenditure growth : reassessing the threat of ageing," 2007.

[18] P. H. M. Van Baal *et al.*, "Lifetime medical costs of obesity: Prevention no cure for increasing health expenditure," *PLoS Med.*, vol. 5, no. 2, pp. 0242–0249, 2008.

[19] E. Gummerson and D. Schneider, "Eat, drink, man, woman: Gender, income share and household expenditure in South Africa," *Soc. Forces*, vol. 91, no. 3, pp. 813–836, 2013.

[20] A. Bilgic and S. T. Yen, "Household alcohol and tobacco expenditures in Turkey: A sample-selection system approach," *Contemp. Econ. Policy*, vol. 33, no. 3, pp. 571–585, 2015.

[21] K. B. Giang, H. Van Minh, and P. Allebeck, "Alcohol consumption and household expenditure on alcohol in a rural district in Vietnam.," *Glob. Health Action*, vol. 6, p. 18937, 2013.

[22] G. Gordon-Strachan *et al.*, "Richer but fatter: The unintended consequences of microcredit financing on household health and expenditure in Jamaica," *Trop. Med. Int. Heal.*, vol. 20, no. 1, pp. 67–76, 2015.

[23] H. Jiang, M. Livingston, and R. Room, "How financial difficulties interplay with expenditures on alcohol: Australian experience.," *J. Public Heal.*, vol. 23, no. 5, pp. 267–276, 2015.

Hsin-Yen Yen is a RN and PhD student in Graduate Institute of Sport, Leisure, and Hospitality Management, National Taiwan Normal University. He also had an education master degree in Graduate Institute of Sport and Leisure Management, National Taiwan Normal University, Taiwan. Besides, he also received bachelor degree in science of nursing, Taipei Medical University, Taiwan. His research interests are physical activity, sport therapy, health promotion, chronic diseases, wellness tourism, and data analysis.

Ching Li is a professor. Dr. Li is a professor and chairman of Graduate Institute of Sport, Leisure, and Hospitality Management, National Taiwan Normal University, Taiwan. She had a PhD degree in environmental science, State University of New York, NY, USA. Her research expertise focuses on recreational resource management, environmental planning and assessment for sport facilities, community recreation, leisure behavior, and hiking.

Ping-Feng Hsia is a PhD student and had a master of education degree in Graduate Institute of Sport, Leisure and Hospitality Management, National Taiwan Normal University. She also received bachelor of administration degree in the Department of Sport and Leisure Management, National Taipei University, Taiwan. Her research interests are tourist behavior especially in information search, tourism destination management, and data analysis.