Applying Data Mining Technology on the Using of Traditional Chinese Medicine in Taiwan: An Application of C5.0 Decision Tree

Yi-Horng Lai

Department of Health Care Administration, Oriental Institute of Technology
New Taipei City, Taiwan
Email: FL006 {at} mail.oit.edu.tw

ABSTRACT— *This research employed the complete datasets of Traditional Chinese medicine (TCM) outpatient reimbursement claims from 2005 to 2007 to analyze the using of TCM, the characteristics of TCM patients, and the disease categories that were treated by TCM in Taiwan. With the result of this study, female use TCM more frequently than male, is consistent with previous reports from western countries. The reasons for this female predominance were not fully elucidated in previous reports. It was suggested that independent females, or females of good social status, had higher expectations of or belief in TCM in respect of postpartum conditions, menopause and chronic diseases. The age distribution of TCM users peaked in the 20-29 group, followed by the 10-19 group and 31-39 group. Most TCM visits were to private TCM clinics, followed by the private TCM hospitals. According to the results, the most common reasons for TCM visits were Kuru (460), Cough (786.2), Allergic rhinitis cause unspecified (477.9), Lumbago (724.2), Headache (784.0), Myalgia and myositis, unspecified (729.1), Constipation (564.0), Other sleep disturbances (780.59), Sleep disturbances, unspecified (780.50), and Dyspepsia and other specified disorders of function of stomach (536.8). TCM was popular in the Chinese population. More and more subjects used TCM at least once during the 3-year study period. TCM, like western medicine, was commonly used by the Chinese population for problems and diseases of major human organ systems. This study provides information about the use frequencies of TCM and disease categories treated by TCM, which should be useful for health policy makers and for those who consider the integration of Chinese and Western medicine.*

Keywords—Traditional Chinese medicine, data mining, C5.0 decision tree

---

1. INTRODUCTION

Traditional Chinese medicine (TCM) was an important topic of complementary and alternative medicine in Western opinion [9]. Current TCM practices can be traced back more than 2000 years. TCM was still commonly used by the Chinese [3]. In Taiwan, not until the 1980s did several researchers start to research issues relevant to TCM, using sampling surveys or studies with small sample sizes [4]. There has been no large-scale investigation of the use of TCM among Chinese people worldwide, now.

Importance in complementary and alternative medicine had increased substantially in western countries during the past decade [1]. Patients and their families seem to have sought their health practitioners' opinions about various complementary and alternative medicine modalities more frequently [2]. Many studies have demonstrated dramatic increases in the use of, and expenditure on, complementary and alternative medicine in the Western opinion [2]. However, most of the prevalent studies of complementary and alternative medicine use were based primarily on questionnaire surveys, telephone interviews or collecting data from insurance claims, and the sample sizes generally were small.

In Taiwan, the National Health Insurance (NHI) was started in 1995 and covers nearly all inhabitants (21,653,555 beneficiaries at the end of 2001) [5]. The use of TCM has been reimbursed by the NHI since 1996. Taiwanese were free to choose Western medicine or TCM, and were allowed to visit either public or private medical facilities. Because all claims data are available to researches in electronic form, it could be conduct a study of TCM use among the Chinese population in Taiwan.

The aim of this study was to conduct a nation-wide survey in order to investigate the using of TCM, the characteristics of TCM users, and the medical conditions for which Taiwanese people most commonly use TCM, by analyzing the NHI claims data from 1997 to 2010. TCM provided by the NHI included Chinese herbal remedies,

acupuncture and traumatology manipulative therapy.

2. METHODOLOGY

2.1 Data Sources

The NHI program was initiated in Taiwan since 1995 and covers nearly all inhabitants. In 1999, the Bureau of NHI began to release all claims data in electronic form to the public under the National Health Insurance Research Database (NHIRD) project. The structure of the claim files is described in detail on the NHIRD website and in other publications [4].

It could be obtain the complete TCM claim datasets from the NHIRD in Taiwan. The datasets contained only the visit files, including dates, medical care facilities and specialties, patients' genders, dates of birth, and the three major diagnoses coded in the International Classification of Disease, 9th Revision, Clinical Modification (ICD-9-CM) format [6]. To protect privacy, the data on patient identities and institutions had been scrambled cryptographically.

These visit files represented all the TCM outpatient activities within the NHI from 2005 to 2007. Insurance benefits were available for TCM that included Chinese herbal remedies, acupuncture and traumatology manipulative therapy, especially for joint dislocation. In Taiwan, TCM is reimbursed by NHI only in ambulatory clinics, not for inpatient care. In addition, only licensed TCM physicians qualify for reimbursement from the NHI.

2.2 Study Design

Although the concept of disease entities in TCM was quite different from that in Western medicine, TCM physicians are requested to follow the standard diagnoses according to the ICD-9-CM coding system when claiming reimbursement. Common diagnostic groups for TCM visits were categorized according to the reclassification of primary ICD-9-CM codes for use in the National Ambulatory Medical Care Survey and National Hospital Ambulatory Medical Care Survey data in the United States.

To calculate patients' ages in relation to the 3-year use frequency of TCM from 2005 to 2007, December 31, 2007 was taken as the index of subtrahend. The denominator was the number of people who were insured during this 3-year period.

In order to compare the average numbers of visits between TCM and Western (allopathic) medicine, it could be obtained the total number of ambulatory visits to Western medicine from the website of Department of Health, Taiwan. In addition, we obtained the sampling claim datasets for ambulatory care visits at Western medicine clinics in order to compare the top ten disease categories between TCM and Western medicine visits. The sampling was random and visit-based but was separated monthly to eliminate possible seasonal variations. According to the NHIRD, these sampled files were representative of all utilization within the NHI in Taiwan.

2.3 Methodology

Decision tree learning was a method commonly used in data mining. The goal was to create a model that predicts the value of a target variable based on several input variables. Each interior node corresponds to one of the input variables; there were edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

A decision tree was a simple representation for classifying examples. Decision tree learning was one of the most successful techniques for supervised classification learning. For this section, assume that all of the features have finite discrete domains, and there was a single target feature called the classification. Each element of the domain of the classification is called a class. A decision tree or a classification tree was a tree in which each internal node was labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature. Each leaf of the tree was labeled with a class or a probability distribution over the classes.

A tree can be learned by splitting the source set into subsets based on an attribute value test. This process was repeated on each derived subset in a recursive manner called recursive partitioning. The recursion was completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions.

In data mining, decision trees could be described also as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data.

Data comes in records of the form:

$(x,Y)=(x_1, x_2, x_3, \ldots, x_k, Y)$

The dependent variable, Y, is the target variable that we are trying to understand, classify or generalize. The vector x is composed of the input variables, x_1, x_2, x_3 etc., that are used for that task.

C5.0 decision tree was developed by Quinlan. [10] To compute Gini impurity for a set of items, suppose i takes on values in {1, 2, ..., m}, and let fi be the fraction of items labeled with value i in the set.

$$I_G(f) = \sum_{i=1}^{m} f_i(1 - f_i) = \sum_{i=1}^{m} (f_i - f_i^{2}) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} f_i^{2} = 1 - \sum_{i=1}^{m} f_i^{2}$$

Used by the C5.0 tree-generation algorithms. Information gain is based on the concept of entropy from information theory.

$$I_E(f) = \sum_{i=1}^{m} f_i \log_2 f_i$$

2.4 Software for Data Mining

IBM SPSS Modeler 14.1 was the main software used for data linkage and processing. Descriptive data, including frequencies, percentage and means, are presented.

3. RESULTS

Among the 1089885 valid beneficiaries of the NHI program at the end of 2007, 16153 had used TCM during the year, with a total of 1547708 visits. The annual number and percentage of TCM users steadily increased from 2005 to 2007 (as Table 1). The age distribution of the TCM users peaked in the 20-29, followed by the 10-19 and 30-39 (as Table 2), while the age distribution for visit counts showed a peak in the 40-49 followed by the 30-39 and 20-29.

Table 1: Patient use and visit counts of traditional Chinese medicine (TCM) within national health insurance (NHI) from 2005 to 2007 in Taiwan

Year	Valid beneficiaries within NHI	Total No	Subjects using TCM		New patient	Total visits
			Female	*Male*		
2005	370211	113832	65860	47972	113832	588079
2006	362007	110715	64426	46289	43650	571634
2007	357667	113244	66145	47099	31121	599603
Total	1089885	337791	196431	141360	188603	1759316

Table 2: age-specific usage frequency of traditional Chinese medicine (TCM) during the 3-year period from 2005 to 2007 in Taiwan

Age (years)	Number of total population	Number of subjects using TCM	(%)	Number of TCM visits
≤9	43977	13607	30.94	108557
10–19	54409	26006	47.80	174824
20–29	64656	34869	53.93	258758
30–39	64046	33418	52.18	322126
40–49	62566	32775	52.38	354501
50–59	46130	24135	52.32	264651
60–69	26232	12871	49.07	144838
70–79	19375	8348	43.09	100455
≥80	8318	2574	30.94	30606

Base on the age group and sex, the result of C5.0 decision tree was as Figure 1. Predictor importance of age group was .66 and sex was .34 (as Figure 2).
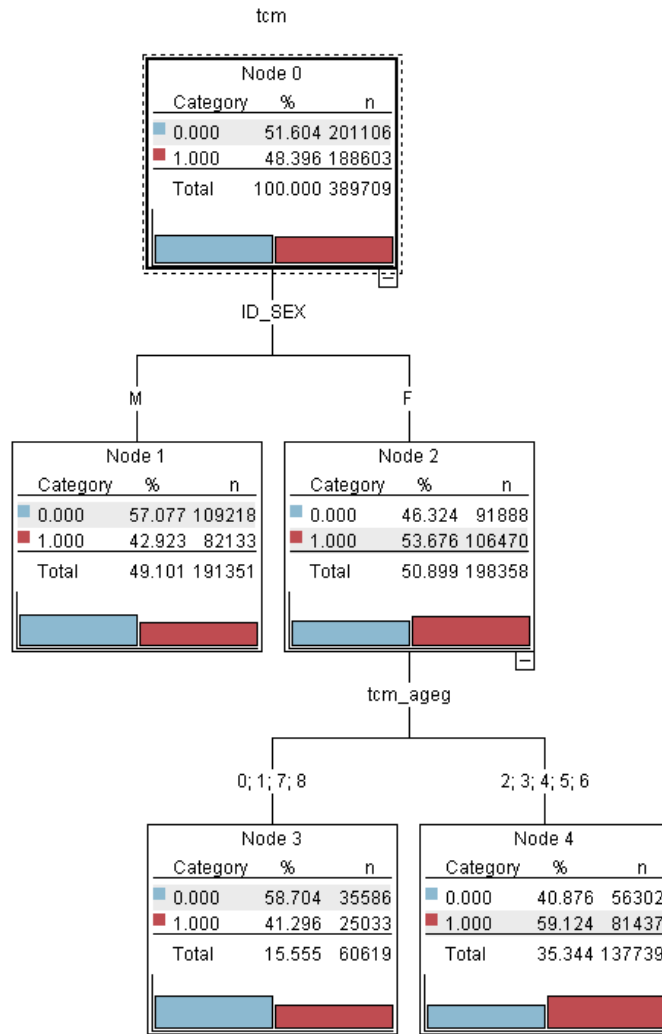
Figure 1: The Result Of C5.0 Decision Tree
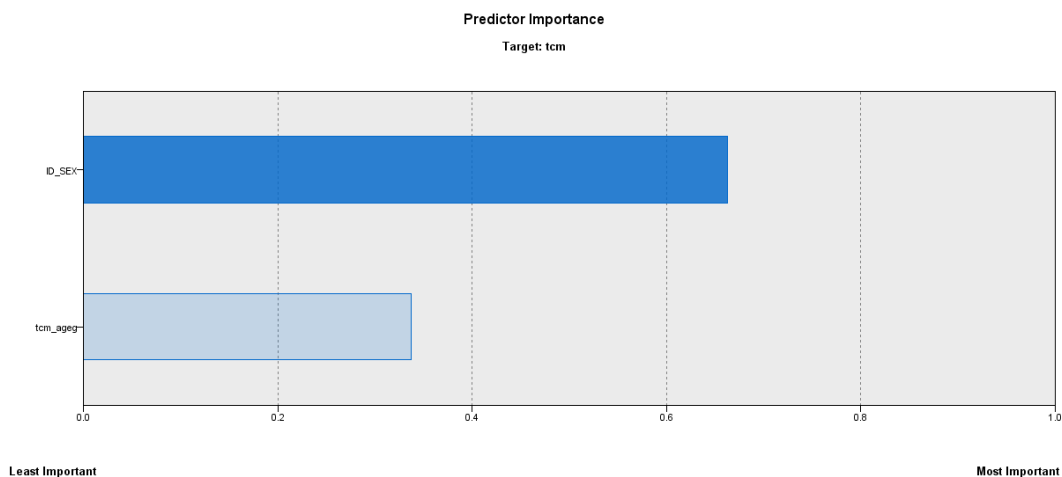

Figure 2: Predictor Importance

Most of the TCM visits identified in the study were performed in private TCM clinics (11586971), followed by private TCM hospitals (126277), public TCM hospitals (45665) and public TCM clinics (403). Visits to private TCM hospitals decreased yearly, while visits to private TCM clinics, public TCM hospitals and others increased (as Table 3).

Table 3: service volume of traditional Chinese medicine (TCM) by facility type from 2005 to 2007 in Taiwan

Year	Public TCM hospital	Private TCM hospital	Public TCM clinics	Private TCM clinics	Total
2005	15202	44002	145	528730	588079
2006	14802	41113	157	515562	571634
2007	15661	41162	101	542679	599603
Total	45665	126277	403	1586971	1759316

Among the 1947696 TCM visits, each of them had one, two, or three clinical diagnoses according to the ICD-9-CM coding system. The top ten diseases for TCM visits were Kuru (7.64%), Cough (5.04%), Allergic rhinitis cause unspecified (3.50%), Lumbago (3.07%), Headache 2.99%), Myalgia and myositis, unspecified (2.52%), Constipation (2.43%), Other sleep disturbances (2.02%), Sleep disturbances, unspecified (1.82%), and Dyspepsia and other specified disorders of function of stomach (1.80%) (as Table 4).

Table 4: The top 10 major disease categories for traditional Chinese medicine visits from 2005 to 2007 in Taiwan

Disease	ICD-9-CM code	Number of visits	(%)
Kuru	460	148835	7.64
Cough	786.2	98216	5.04
Allergic rhinitis cause unspecified	477.9	68255	3.50
Lumbago	724.2	59877	3.07
Headache	784.0	58141	2.99
Myalgia and myositis, unspecified	729.1	49021	2.52
Constipation	564.0	47358	2.43
Other sleep disturbances	780.59	39388	2.02
Sleep disturbances, unspecified	780.50	35353	1.82
Dyspepsia and other specified disorders of function of stomach	536.8	35082	1.80
Others		1308170	67.16
Total		1947696	100.00

Furthermore, it could be analyzed the percentage distribution of major disease categories for TCM visits by age (as Table 5). The results show that Kuru (460), Allergic rhinitis cause unspecified (477.9), Dyspepsia and other specified disorders of function of stomach (536.8), Myalgia and myositis, unspecified (729.1), and Other sleep disturbances (780.59) were higher in the 31-39 groups. The results show that Constipation (564.0) was higher in the 21-29 groups. The results show that Lumbago (724.2), Headache (784), and Flatulence, eructation, and gas pain (787.3) were higher in the 41-49 groups. The results show that Cough (786.2) was higher in the 41-49 groups.

It could be found no significant differences between males and females in the percentage distributions of the commonest disease categories for TCM visits (Table 6). However, female subjects visited TCM for Kuru (460), Dyspepsia and other specified disorders of function of stomach (536.8), Constipation (564.0), Lumbago (724.2), Myalgia and myositis, unspecified (729.1), Other sleep disturbances (780.59), Headache (784), Cough (786.2), and Flatulence, eructation, and gas pain (787.3) more frequently than males.

It also be compared the percentage distribution of major disease categories for TCM visits among different locations and the results revealed that Kuru (460), Cough (786.2), Lumbago (724.2), Headache (784.0), Myalgia and myositis, unspecified (729.1), Allergic rhinitis cause unspecified (477.9), Constipation (564.0), Dyspepsia and other specified disorders of function of stomach (536.8), Flatulence, eructation, and gas pain (787.3), and Other sleep disturbances (780.59) were more commonly seen in clinics than in hospitals (Table 7).

Table 5: percentage distribution of diseases categories for traditional Chinese medicine visits by different age groups,2005–2007, in Taiwan

ICD-9	-9	10-19	21-29	31-39	41-49	51-59	61-69	71-79	81-

code									
460	23698	21015	18670	26264	25677	17753	8666	5514	1578
477.9	15580	17464	9885	10474	7490	4155	1624	1254	329
536.8	3677	4038	4548	5987	6849	4775	2664	1990	554
564.0	1882	3987	11269	9631	8312	5366	2526	3048	1337
724.2	126	2174	7575	10474	13344	11716	7056	5426	1986
729.1	449	3039	7914	9576	11293	8683	4364	2814	889
780.50	193	783	4011	6840	8635	8008	4136	1996	751
780.59	123	1111	4925	8549	10529	8028	3562	2092	469
784.0	580	3067	7465	13021	15396	9826	4650	3299	837
786.2	19179	11497	8523	14819	15954	12508	8035	6077	1624
Total	65487	68175	84785	115635	123479	90818	47283	33510	10354

Table 6: number of visits and percentage distribution of diseases categories for traditional Chinese medicine visits by gender, 2005-2007, in Taiwan

ICD-9 code	*Female*	*Male*	*Total*
460	89681	59154	148835
477.9	33667	34588	68255
536.8	20117	14965	35082
564.0	38120	9238	47358
724.2	31729	28148	59877
729.1	29423	19598	49021
780.50	22871	12482	35353
780.59	25105	14283	39388
784.0	42928	15213	58141
786.2	55394	42822	98216
Total	389035	250491	639526

Table 7: percentage distributions of diseases categories for traditional Chinese medicine visits by location, 2005-2007, in Taiwan

ICD-9 code	*Public TCM hospital*	*Private TCM hospital*	*Public TCM clinics*	*Private TCM clinics*	*Others*
460	1346	4852	21	142616	148835
477.9	4875	11537	14	51829	68255
536.8	1868	2298	4	30912	35082
564.0	2289	4664	23	40382	47358
724.2	2064	4659	22	53132	59877
729.1	2142	6020	3	40856	49021
780.50	3125	2666	16	29546	35353
780.59	520	1175	0	37693	39388
784.0	1064	2918	13	54146	58141
786.2	2119	4407	38	91652	98216
Total	21412	45196	154	572764	639526

4. DISCUSSION AND CONCLUSION

After all, this study is the first extensive survey of TCM use in Chinese society. Only with the help of a computerized insurance reimbursement database could such a large-scale TCM utilization study feasibly be analyzed. Previous studies from western countries on the frequency and characteristics of TCM use have mainly consisted of surveys of clinic attendees, telephone interviews, written surveys, household interviews, and hospital and private clinic surveys; and the sample sizes have been limited. In addition, the use of TCM in western countries is usually not covered by insurance [7]. Thus, the survey results might be affected by the socio-economic status of the subjects [8]. Fortunately, TCM is reimbursed by NHI in Taiwan, so the study would seem to be less biased.

The use of TCM in western countries has increased dramatically in recent years [9]. It goes without saying that TCM had been commonly used in Asian countries, especially in the Chinese population, for centuries [4]. Owing to the

different definitions of complementary and alternative medicine, the types of complementary and alternative medicine surveyed, survey methodologies and types of complementary and alternative medicine reimbursed by insurance, it was difficult to compare the use frequency of TCM among countries [11]. According to the results, there was a steady increase in the annual number of TCM users in Taiwan between 1997 and 2010; this does not include folk medicine, which is not reimbursable by insurance. Chinese people believe that Western medicine may react faster to the target but also causes more adverse side effects, while TCM reacts slowly but is subtle and safe [8,9,10]. Furthermore, the insurance coverage for TCM visits might also play a significant role [8].

With the result of this study, female use TCM more frequently than male, is consistent with previous reports from western countries. The reasons for this female predominance were not fully elucidated in previous reports. It was suggested that independent females, or females of good social status, had higher expectations of or belief in TCM in respect of postpartum conditions, menopause and chronic diseases.

The age distribution of TCM users peaked in the 20-29 group, followed by the 10-19 group and 31-39 group. Most TCM visits were to private TCM clinics, followed by the private TCM hospitals.

According to the results, the most common reasons for TCM visits were Kuru (460), Cough (786.2), Allergic rhinitis cause unspecified (477.9), Lumbago (724.2), Headache (784.0), Myalgia and myositis, unspecified (729.1), Constipation (564.0), Other sleep disturbances (780.59), Sleep disturbances, unspecified (780.50), and Dyspepsia and other specified disorders of function of stomach (536.8).

TCM was popular in the Chinese population. More and more subjects used TCM at least once during the 3-year study period. TCM, like western medicine, was commonly used by the Chinese population for problems and diseases of major human organ systems. This study provides information about the use frequencies of TCM and disease categories treated by TCM, which should be useful for health policy makers and for those who consider the integration of Chinese and Western medicine.

5. ACKNOWLEDGEMENT

6. REFERENCES

[1] Ahmad, A.W. "Determinants of complementary alternative medicine (CAM) use," Complementary Therapies in Medicine, vol. 12, pp. 99-111, 2004.

[2] Burg, M.A., Hatch, R.L., & Neims, A.H. "Lifetime use of alternative therapy: A study of Florida residents," Southern Medical Journal, vol. 91, pp. 1126-1131, 1998.

[3] Chen, F.P., Kung, Y.Y., Chen, T.J., & Hwang, S.J.: "Demographics and patterns of acupuncture use in the Chinese population: The Taiwan experience," Journal of Alternative and Complementary Medicine, vol. 12, pp. 379-387, 2006.

[4] Lee, C.H., Chou, Y.J., Chen, L.S., & Chang, H.J. "Utilization of ambulatory Chinese medical services under the National health Insurance in Taiwan," Taiwan Journal of Public Health, vol. 23, pp. 100-107, 2004.

[5] Cheng, S.H., & Chiang, T.L. "The effect of universal health insurance on health care utilization in Taiwan: results from a natural experiment," JAMA, vol. 278, pp. 89-93, 1997.

[6] Liu, J.Y., Chen, T.J., & Hwang, S.J. "Concomitant prescription of nonsteroidal anti-inflammatory drugs and antacids in the outpatient setting of a medical center in Taiwan: A prescription database study," European Journal of Clinical Pharmacology, vol. 57, pp. 505-508, 2001.

[7] Cleary-Guida, M.B., Okvat, H.A., Oz, M.C., & Ting, W. "A regional survey of health insurance coverage for complementary and alternative medicine: Current status and future ramifications," Journal of Alternative and Complementary Medicine, vol. 7, pp. 269-273, 2001.

[8] Wolsko, P.M., Eisenberg, D.M., Davis, R.B., Ettner, S.L., & Philips, R.S. "Insurance coverage, medical conditions, and visits to alternative medicine providers- results of a national survey," Archives of Internal Medicine, vol. 162, pp. 281-287, 2002.

[9] Eisenberg, D.M., Davis, R.B., Ettner, S.L., Appel, S., Wilkey, S., Van Rompay, M., & Kessler, R.C. "Trends in alternative medicine use in the United States, 1990–1997: results of a follow-up national survey," JAMA, vol. 280:1569-1575, 1998.

[10] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. lassification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8, 1984.

[11] Zollman, C., & Vickers, A. "ABC of complementary medicine: Users and practitioners of complementary medicine," BMJ, vol. 319, pp. 836-838, 1999.