ƒ ‹ F | Œfia‹ ˙'e›ß• |–M^D‹ ª¤s›p e ƒ¤"G‡ł§i

¥H¶¥…hƒ¡´O¶ ‹ ª¤sƒh›«¥~¶¸

›p e^ §¡G §O«‹ ›p e

›p e‰ ¸s„¡GNSC91-2213-E-038-004-

ıƒ · ¶¡¡G91 ƒ~08 ⁄ °01 ⁄ Ø 92 ƒ~07 ⁄ °31 ⁄Ø

ıƒ ‡ ƒ ¡G» O ¥´ ˙⁄j ˙´ ˙, ŒT‹ ª¤s'

›p e¥ D ø⁄¡HG‰–¥H⁄fl

‡ł§i^ «‹¡G" °†‡ł§i

‡ B ‡⁄ƒ¡¡G¥»›p efl A¸–M§ Q ' ¸ ¤ ¥ L · …z1 ƒ†«£ Æv¥Ai ⁄ ¶}‹ d,

⁄⁄ ¥` Œ 92 ƒ~12 ⁄ °26 ⁄Ø

¥H¶¥..hƒ¡´O¶ ⁄ "R« ª¤s´ ·,Œfi˘fiw⁄Sƒh›«¥~¶¸

Hierarchical Clustering of Multiple Traumas in Medical Database

%–¥H⁄fl¸... B Þ'¸M‡ fl •"¸N Š ı ″–¸M fl ¸ Ł ¯

Chiang, I-Jen [ab]; Liu, Charles C.H.[abc]; Chen, Ray-Jade [ad]; Li, Yu-Chuan [a]; Wong, Chau-Min [b]

[a] ¥ x ¥ _ ´ · ⁄ j · ´ · ¸[b]Œ¥xŸW ª¤s' ´ ⁄ u{ · ‹ ªs'

[c] Œfiı´ · ⁄ ⁄ ¥~‹ [d]" Ł ´ · ⁄ ⁄ ¥~‹¸¥~¶¸‹

¥ D ø⁄H % – M⁄fl › p ªs:N¸SC 91–2213–E–038–004

⁄@¡B⁄⁄ "K›n

´ ·¸fŒ˘ – · ‹2001 ƒ~¥x˘Wƒ a ˘ • « O Š % ˜ ¥X¸Œfi«fiw (NHIRD) ¤ ‹ ªs¤ ¥¸ « Æ " " … " ø ^ Þ ¥¸Œ¥›¸¥¸¶¶‹¸ ¤¡A – · Q ¶ ¥ … h ƒ ¡ ´ ¶ ⁄j«"¸Rœfi˘„fi⁄W ¤ Š U´ ‹ª¤ş´ | " ‡⁄Š›«° » PA¥ ' ¡C

¶ ¥ … h ƒ ¡ ´ O ¶ ⁄ " Rªƒ‡"ª†{ O¶ ⁄ I u ´ I ¡ A · £ ¤ – M fiƒ›‹ ª$¥ "„" ⁄ ¤ ÷ ¸Œfi ˘ † " ç C¥H Matlab „ Œ Š»P ¸Œfi ˘ fi w ` p †¡Œfi£¥˝ › x " ‰u⁄W ¸R (OLAP) ' » ⁄ Šß«Œfi ˘ " Ą¤⁄Œ¸B¸Œfi˘ ⁄ " R ¡ C „',Œfi«Ofi w ¤ ‰^¥Œ›Ł"„ ICD ¶ É_¸ŒT" " ª † z ¡ A Š › ⁄ ⁄J {Š ⁄ ¤ ‡¸ (CCS) Š@¸Œfi ˘ " " " Þ ª†z¡C

Š › ‡¤¥D^ D ⁄ W ¶†–ƒb ‹ª¤ş o † { ‴† ,·" (1)¥H´ O ¶ – ˘ Š «Æ⁄Š „ ˘ fi†¥(x˘Wƒ a ƒh¥›«¶¸⁄ Š ¥ ‹ – ¡ Þ ¡A¥¡ · £ ¤ ‹ F ƒ ¤ ¥W»Þ(2)¥Œ⁄Š¥¶ ⁄ "R ƒ fl¯´ | ‡¿¶Š¸ƒ | fi "„"˘ƒ›fiƒP Ł ^«‹ ¡Œ ƒ ⁄ , s ¥ i ¤ Š U o–⁄H ƒP%Ł,' ş"A'flÞ Š " ° †"‰Þ¸ " R Ą'¥Þ¸o† { † Š ±"„"flÞ ¤ '¸´ ¦C

^ ` ¡G¸Œfi ˘ ÷ , ´ O ¶ ⁄ ,"R ·,fŒ˘ fiw,¥ †Ł¸Œfi fiw, ¥ ¥` «Œ‹ ª ¤,fŒfiw, ¥~¶¸, ¿S¶¸

ABSTRACT

Medical data mining becomes more mandatory after the release of national healthcare insurance research database (NHIRD) in 2001. In multiple trauma injuries, our research demonstrated the feasibility and benefits of the hierarchical clustering methodology in medical research and in hospital administration.

The process of clustering could be explicit shown in hierarchical cluster analysis. The domain experts could elucidate the structure of data in interactive setting. We implement in Matlab the realtime linkage with the data sources, and propose the methodology of dimensional table generation for further online analytical processing (OLAP) in a data warehouse. For the widely varied coding policy in the International Classification of Disease (ICD) diagnosis fields, we introduced Clinical Classification System (CCS) for more medically relevant data preparation.

The medical findings were illustrated in details in two topics – (1) Visualization of the distribution and correlation of multiple trauma in Taiwan, which could provide abundant information for research needs from clinical and administrative fields; (2) To find the major treatment patterns of hospitalized burn patients in various hospitals, to facilitate sophisticated statistical analysis of more homogeneous patient populations after clustering, or as the surveillance of normal pattern and to find patient or hospital/doctor outliers.

Keywords: Data mining, Cluster analysis, Medical database, Claim database, National Healthcare Insurance Research Database (NHIRD), Trauma, Burn injury.

BACKGROUND:

The release of National Health Insurance Research Database (NHIRD) in 2001 brought a new challenge to the medical informatics society in Taiwan [11]. In addition to its huge size and the need of data warehouse for management, the issue of data quality and complexity, as mixture of various

clinical severity and indications from heterogeneous sources, rendered more difficulties on the researchers from clinical medicine or hospital administration.

To find the risk factors contributing to the clinical outcome is always one of the major aims of clinical research. The interplay of co-mobidities and complications is especially important in the retrospective insurance claim database [7]. In our preliminary experience in NHRID, to add more clinical details, and to focus on more homogeneous patient subpopulations was the key factor of yielding data mining in the heterogeneous medical datasets [10][13].

Hierarchical clustering (HC) is recently revisited vigorously for phylogenomics study and the genetic expression pattern of microarray [3][5][8]. We use the agglomerative algorithm (HAC) for the same reason as in bioinformatics, in order to allow intervention by domain knowledge in data rescaling and in the intermediate stages of clustering, to explore the unknown pattern of diagnosis and treatment pattern in the dataset of 96% coverage of Taiwan population.

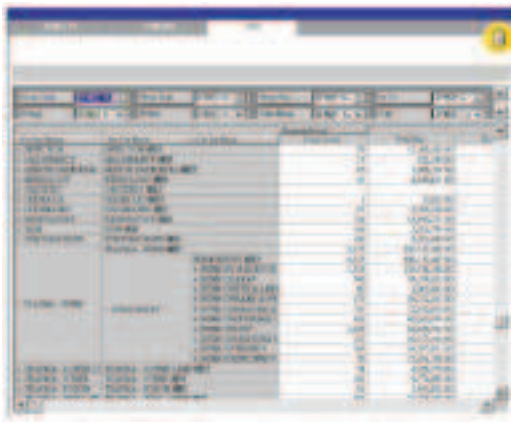

Fig.1. OLAP display of hierarchical ICD categories of disease and statistics of subgroup

Table 2-1. Sorted table of cluster information, at 200 cluster cutoff level, only the clusters with case number larger than 5 were shown. (Data for SB clustering in RESULT section 4.2.1).

c00 5	c01 0	c01 5	c02 0	c05 0	c07 5	c10 0	c12 5	c15 0	c17 5	c20 0	CaseNo	BURN CD	DEBRI DE	STSG	VE NT
5	3	1	11	30	7	5	11	51	19	60	23	9,097	1,393	0	0
5	3	1	11	30	7	5	11	51	20	61	47	6,672	932	0	0
5	3	1	11	30	7	5	12	52	38	76	9	4,900	4,580	0	0
5	3	1	11	30	7	5	12	52	42	15	6	4,737	8,930	0	0
5	3	1	11	30	7	5	12	52	42	16	9	5,009	12,501	0	0
5	3	1	11	30	7	6	45	82	1	47	12	11,638	469	0	0
5	3	1	11	30	7	29	61	40	16	59	309	395	1,850	0	0
5	3	1	11	30	7	29	61	40	17	19	45	1,677	4,173	0	0
5	3	1	11	30	7	29	61	40	17	20	16	3,001	6,740	0	0
5	3	1	11	30	7	29	61	40	17	36	34	274	6,890	0	0
5	3	1	11	30	7	29	61	40	18	21	43	4,273	674	0	0
5	3	1	11	30	7	29	61	40	18	22	113	2,429	762	0	0
5	3	1	11	30	7	33	64	38	76	17	19	49	12,247	0	0
5	3	1	11	30	13	42	71	13	4	50	8	7,289	16,299	0	0
5	3	1	11	30	13	42	71	13	5	51	7	9,240	18,407	0	0
5	3	1	11	30	30	9	34	72	28	5	17	1,672	1,888	5,099	0
5	3	1	11	30	30	9	34	72	28	6	32	61	2,119	5,113	0
5	3	1	11	30	30	9	34	72	29	68	14	560	5,392	5,104	0
5	3	1	11	30	30	37	16	56	89	118	6	0	2,379	8,084	0

Table.1. Dimensional table for OLAP service, generated directly from the clustering results

MATERIALS AND METHODS:

Data preparation and warehousing

The trauma subsets of NHIRD from 1997 to 2001 were pooled into a Oracle 8i data warehouse. Another copy was managed in a Microsoft SQL 2000 for comparison. About 20G of disk space were required for each copy, and the OLAP manipulation might need additional 10 to 40 G.

The hospitalized trauma patients with more than two diagnoses were extracted, totally 1,319,176 cases.

To smooth out the various coding policies of different hospitals of similar clinical conditions, we introduced AHRQ clinical classification system (CCS) ICD remapping [2].

To study the treatment pattern of burn trauma, 1519 cases with information about area of injury were extracted from the 2860 cases in the 1/20 sampled datasets with clinical details.

Clustering programs

The TreeView and Hierarchical Cluster

Explorer (HCE) available in the public domain during the bioinformatics hype were surveyed [4][6].

Mathwork Matlab 6.1, with the Database and Statistical toolboxes, was used for data connection and clustering. The cases were sorted in clusters, and the dimensional table was then generated, in regular user-specified intervals (based on the number of clusters or criteria of similarity), then delivered to the OLAP service. An example were given in Table 1 and Figure 1.

Exploratory data analysis was performed in the java-based IndexMiner data mining package, developed by the senior author [1].

aVarious combinations of data rescaling, normalization, linkage and similarity function were tried [3][8].

RESULT & DISCUSSION:

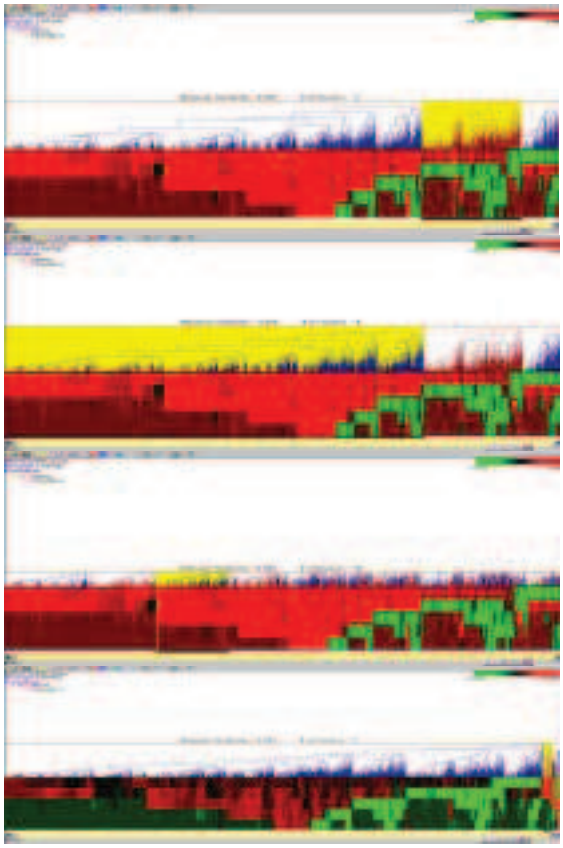

Fig.2a,b,c.. HAC of multiple trauma in Taiwan. Red designated for various trauma injuries; green, for associated nontraumatic diseases. Fig.2d, burn cases highlighted after data rescaling. (by HCE)

Distribution of multiple trauma

The CCS numerical codes were clustered after rescaling the trauma codes (from 16.1 to 16.9) to between 1 and 9 (red), and recoding of the other non-traumatic disease to their negative values (green) with single linkage (Fig.2). If the criterion of similarity was set too high, only the clusters with much variable disease combination were shown (Fig.2a). The lower, and more homogeneous, clusters need change of the threshold for separation, comparing Fig.2c with Fig.2b. The interactive nature supported the need of customizable views in a data warehouse.

To show the burn injury of particular interest, the burn diagnosis was recoded to 18, and then the burn clusters with or without associated injuries or diseases were highlighted clearly.

Treatment patterns of burn injury

According to the "vector models" in text mining[9], the most frequently present medical orders or the orders with statistical significance according to users' interest yielded small clusters (Fig.3). Some of the purified patient groups could arouse clinical interest, but the overall pattern were not clear.

After data normalization and "windowing" specific to the clinical interest, the importance of management of extreme values and missing values was stressed.

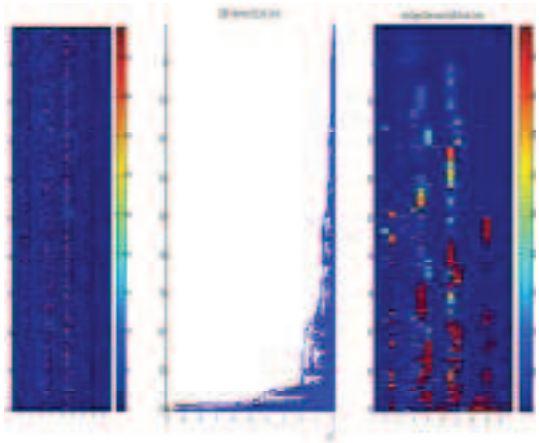

Fig.3a. Unsorted burn data set; Fig.3b,c, the dendrogram and respective clusters after sorting. (by Matlab)

Dressing-change, debridement operations, and skin grafting were finally used as criteria of clustering (Table 2). Six groups of treatment patterns were noted (Fig.4). More heavy-grafting type and heavy-dressing-change type occurred in hospitals with burn specialties. No-grafting-and-low-fee and more-dressing-change-than-debridement styles were favored by hospitals without burn specialties.

Clustering of treatment patterns improved the fit and decreased the errors of regression models for prediction of total medical expenses, and helped the judgment of outliers (Fig.5). After clustering, R square of multiple linear regression was improved from 0.650 to 0.702 and 0.876, and R square of regression trees were improved from 0.697 to 0.790 and 0.924, by demographic factors and cluster information, or by addition of length of stay as an independent variable. (Table 3).

The patterns of clustering could be used

ClaSBWN	Case No	BURN_CD	DEBRIDE	STSG	VENTILATOR	Sd BURN_CD	Sd DEBRIDE	Sd STSG	Sd VENTILATOR
Outlier	27	16,821	46,476	4,724	17,333	14,562	44,460	8,563	17,352
hGhCD	99	15,281	34,811	10,686	3,436	14,241	28,411	6,982	11,778
lGhCD	26	10,829	10,653	4,987	208	11,923	6,240	502	1,859
lGlCD	82	1,165	3,179	5,586	8	1,844	2,597	1,146	0
nGhC	119	11,311	1,673	0	8	7,964	2,059	0	0
nGhD	73	5,731	17,575	0	360	4,259	7,074	0	1,419
nGlCD	373	1,295	1,787	13	44	1,364	2,100	186	423

Table 2. Clustering center of the 4 attributes.

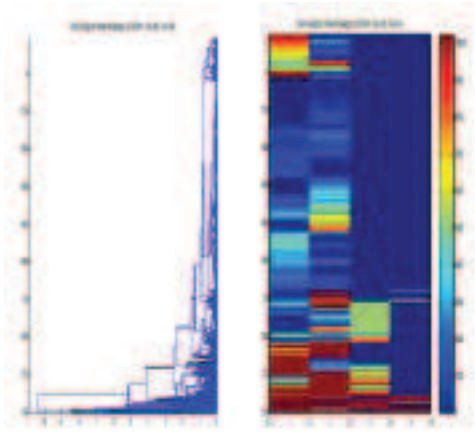

Fig.4. Six major treatment pattern after regrouping of attributes to 1) Dressing-change, 2) debridement operatDressing-change, debridement operations, and 3), 4) two kinds of skin grafting.

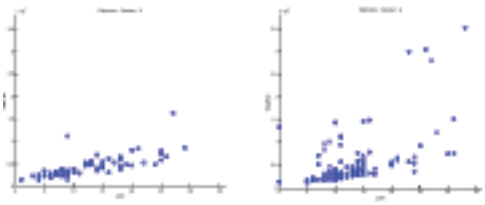

Fig. 5. Hospitalization fee versus days of hospitalization, cleaner data in more homogeneous patient population after clustering.

Table 3-9. Prediction of TotalFee by demographic factors without or with the help of clustering.

		Correlation coefficient	R square	Mean absolute error	Root mean absolute error
Prediction by demographic factors only	Multiple linear regression	0.806	0.650	56,193	95,890
	M5 Regression tree	0.835	0.697	44,315	97,299
By demographic and cluster information	Multiple linear regression	0.838	0.702	46,489	88,438
	M5 Regression tree	0.889	0.790	30,256	74,156
With days of burn ward admission	Multiple linear regression	0.936	0.876	27,599	56,962
	M5 Regression tree	0.961	0.924	16,491	44,801

Table 3-10. Prediction of TotalFee in each clusters.

		Multiple linear regression				M5 Regression tree			
	Case No	Correlation coefficient	R square	Mean absolute error	Root mean absolute error	Correlation coefficient	R square	Mean absolute error	Root mean absolute error
hG	99	0.942	0.887	78,497	115,876	0.949	0.901	69,830	108,397
lGhCD	26	0.759	0.576	38,566	53,481	0.796	0.634	31,747	46,755
lGlCD	82	0.793	0.629	10,002	17,452	0.800	0.640	9,823	16,410
nGhC	119	0.902	0.814	15,041	25,386	0.947	0.897	11,753	18,739
nGhD	73	0.547	0.299	37,785	87,281	0.942	0.887	20,446	28,488
nGlCD	373	0.768	0.590	6,262	12,087	0.81	0.651	5,416	11,112
All Cases	873	0.936	0.876	27,599	56,962	0.961	0.924	16,491	44801

Table.3. Benefits of clustering on the regression tree model for prediction of burn outcome.

as reference for hospital administration, for comparison of high-risk and low-risk patient components. More delicate studies could be arose by the homogeneous patient partitions after cluster analysis.

CONCLUSION & FUTURE WORKS:

We demonstrated the feasibility of hierarchical clustering methodology for research multiple trauma in a huge claim database. The intervention of domain knowledge in many processes of analysis and the interactive nature were addressed, which is especially important for participation of domain experts to handle complicated problems in medical informatics.

In the first example, the overall trauma pattern of the five year NHIRD were explored by the numerical values of CCS

codes, the numerical values were for proximity, and their order was only a byproduct. In the future, we should treat the CCS or ICD codes as simply nominal categories, and define newer distance function.

In the beginning of the second example, the vector model of medical orders could be further explored. The "meaning" of the smaller clusters might be understood with the help of other external utility scores. We think the advancements in text mining and information retrieval could contribute to part of the solution.

REFERENCES:

[1] Chiang I.J., Lin T.Y., Index Miner: a data mining system, In Proceedings of the Annual International Computer Software and Applications Conference (COMPSAC'2001), Chicago, IL, 2001, 613-614.

[2] Clinical Classifications Software (CCS), developed by and available at the Agency for Healthcare Research and Quality (AHRQ), http://www.ahcpr.gov/data/hcup/ccsfact.htm.

[3] Eisen JA. Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. Genome Research 1998; 8:163-167.

[4] Eisen MB, Cluster, and TreeView, current version April 17, 2000. Available at http://rana.lbl.gov/EisenSoftware.htm

[5] Everitt B, Landau S. Cluster analysis. 4th ed ed. London, New York: E. Arnold. Halsted Press, 2001.

[6] Seo J, Bshneiderman B, "Interactively Exploring Hierarchical Clustering Results," IEEE Computer, Volume 35, Number 7, pp. 80-86, July 2002. Software available at http://www.cs.umd.edu/hcil/hce/

[7] Iezzoni LI, Foley SM, Daley J, Hughes J, Fisher ES, Heeren T. Comorbidities, complications, and coding bias. Does the number of diagnosis codes matter in predicting in-hospital mortality? JAMA 267: 2197-203, 1992.

[8] Kaufman L. Finding groups in data an introduction to cluster analysis. New York: Wiley, 1990.

[9] Koller, Daphne and Sahami, Mehran. Hierarchically classifying documents using very few words . Proc. of the 14th International Conference on Machine Learning ICML97. pp.170---178.

[10] Lee, Chien-Chih; Liu, Charles C. H.; Li, Yu-Chuan; Chiang, I-Jen, and Lu, Shiu-Yien. Epidemiology of limb replantation from 1996 to 2000 in Taiwan in National Health Insurance database. Proceedings of Annual Meetings of Taiwan Society of Plastic Surgery, 2002; Taipei. 2002.

[11] National Healthcare Insurance Research Database, http://www.nhri.org.tw/nhird/

[12] Scully KW, Schubart JR, Einbinder J. S. Improving Search Results for Diagnoses or Procedures Using an ICD-9-CM Clustering Standard. Proc AMIA Symp 11332000.

[13] Yang I-Chih, Liu, Charles CH, Li Yu-Chuan, Chiang I-Jen, and Lu Shiuh-Yien. Epidemiology of burn hospitalization from 1996 to 2000 in Taiwan in National Health Insurance database. Proceedings of Annual Meetings of Taiwan Society of Plastic Surgery 2002, 2002.