Text Mining on Chinese Herbal Medicine Rule Exploration for Ovarian Cyst

[1]Dan He, [2]Aiping Lu, [2]Miao Jiang, [3]Guang Zheng, [2]Ning Zhao, [2]Minzhi Wang

[1]E-research institute of Shanghai University of Traditional Chinese Medicine, Shanghai, China,
[2]Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing, China,
[3] School of Information Science and Engineering Technology, Lanzhou University, Lanzhou, Gansu, China,
forzhengguang@126.com fm873@126.com lap64067611@126.com

Abstract - Ovarian cyst (OC) is one of the biggest concerns of women around the world. With the increase in the number of cases of OC, it seems like no woman is safe from this dreaded disease. Traditional Chinese Medicine (TCM) has its advantage in OC management, while due to the complexity and opacity; it is hard to clarify the rules of Chinese herbs. Text mining is a useful method to explore the regularity; we put this technology in principle research of Chinese herbal medicine (CHM) and associated it with patterns of TCM in OC treatment. The results we obtained from this study:Fuling, Guizhi, Taoren, Danpi, Chishao are top five herbs frequently used in OC. The pattern of Qi stagnation and blood stasis is the No.1 syndrome, which is highly coincided with the top lists of the herbs. Conclusion: Text mining is a practical technology, which can help with the research field of medicine regularity and assist the physician with clinical decision; the future research shall be benefited from the outcome mined out by this technology.

Key words - text mining; ovarian cyst; medicine regularity

I . Introduction

Ovarian cyst (OC) is one of the biggest health concerns of women around the world. The number of women suffering from this reproductive problem is quite alarming especially those between the ages 30-60. Women from all occupations are not exempted from having OC. Robert T, et.al has done a prevalence rate investigation among the 15,735 women, and found 2,217 (14.1%) had one or more simple cysts detected at their first fully visualized TVU screening [1]. With respect to the risk factors causing OC, some researches have been done. For example, F Parazzini et.al carried out a case-control study between 1984 and 1994. Cases were 225 women aged less than 65 year with a histologically confirmed diagnosis of benign seromucinous OC admitted to a network of obstetrics and gynaecology departments in Milan and the conclusion is: the risk of seromucinous benign ovarian tumors is greater in more educated women and in women with a history of infertility and with long or irregular menstrual cycles [2].

Some studies reported the pathogenesis of OC, repeated attack of chronic pelvic inflammation, environment, endocrine as well as stress, virus etc. are accepted as the reasons. There are various classifications in clinics [3].The symptoms shared by most types of OC are occasional lower abdominal pain with or without lump, irregular monthly period, dysmenorrhea, leukorrhagia, dyspareunia or vomiting etc.

Currently there is no ideal management for OC; operation would be the best choice for the patients who have surgery indications, while for the patients who have not, biomedical therapy and physiotherapy as well as hormone may be the options. However, no matter which either method physician takes, the clinical efficacy is not satisfied or the patient cannot avoid paying a big bill.

TCM has a long history in curing OC, and valuable experience has been reserved all along. In TCM, OC was belonging to the range of “zhengjia”, “jiju” or “changqin” which are disease names often appeared in Gynecological books. All three mentioned diseases refer to the formed mass while “zhengjia”was used for the problem, which was mostly attributed to the stagnant qi, and the “jiju” mostly refers to the mass caused by blood stasis. Clinical management is dependent on the differentiation of syndrome; however, the patterns, symptoms of OC are changeable and different people hold different opinion due to the characteristics of TCM. Hence what most we want to do is concluding the major principle in curing OC and figure out the inner relationship among herbs, symptoms and patterns, which can afford reference for clinical physicians.

Text mining is a kind of computer technology that can induce some principles through extracting meaningful information and knowledge from a large of textual database, now it is widely used in medical research. Here in our article, we tried to apply the algorithm set based on data slicing [4] with the principle of co-occurrence[5]-[8] ，we mined the dataset of literature on OC. Through setting up the co-existed relationship of herbs, symptoms and patterns, we could get the network for each of them, and then find out the medication regularity of OC.

II . Material and methods

A. Data collection

The dataset is downloaded from SinoMed (http://sinomed.cintcm.ac.cn/index.jsp) with the query term of “ovarian cyst” on July 24, 2012. This dataset comprises 6,205 records of literatures on clinical practices and theoretical research on ovarian cyst, both articles and reviews are included. In the dataset, each record is tagged with its unique ID. All those materials are constituted by titles, keywords and abstracts,etc.and they are the major resources for data mining[9].

B. Data filtering

1) CHM in the plain text format: Based on the keyword list of CHMs (both legal names and other popular names are included for calculation), we filtered the CHMs in the plain text format, and then converted all popular names into legal names. All the CHMs are tagged with their unique paper ID. During the filtering process, each CHM with specific ID will only be calculated once no matter how often it appeared in one paper. Through this way, we can get the frequency list for all identified herbs (Fig 1). What's more, based on the unique paper ID, we could also establish the pairs of co-existed CHMs, Coz they existed in same paper simultaneously. And in line with the principle of permutation and combination, any two different herbs can become pairs except the two of same. For instance, Fuling (Poria in Latin), Guizhi (Ramulus Cinnamomi in Latin), and Danpi (Cortex MoutanRadicisin Latin) are listed within one paper, then we can set up the co-existed relationship such as "Fuling-Guizhi", "Fuling-Danpi" and "Guizhi-Danpi". All those paired herbs will become essential foundation for CHM network construction.
2) CHM in herb formula: Besides the herbs, mentioned in plain text format, there are lots of information in medication regularity could be excavated from the herb formula. Through data-mining skills, we can decompose the formula into individual herbs and calculate the appearance frequency according to their specific ID. Since the herbs were in same formula, we can make up co-existed relationship for any two of herbs according to the array combination principle, which are like what we do with the CHMs in plain text format. In addition, this would be another part of foundation constructing the CHM network.

Table I Top 12 CHMs in merged order

Pinyin Name	Latin Name	Order of Merged	Order of Plain	Order of Formula
Fuling	Poria	1	1	1
Guizhi	RamulusCinnamomi	2	2	2
Taoren	Semen Persicae	3	7	4
Danpi	Cortex MoutanRadicis	4	12	3
Chishao	Radix PaoniaeRubra	5		5
Danggui	Radix AngelicaeSinensis	6	3	6
Baishao	Radix Paoniae Alba	7	4	7
Chuanxiong	RhizomaLigusticiChuanxiong	8		8
Dahuang	Radix et RhizomaRhei	9	5	
Baizhu	RhizomaAtractylodisMacrocephalae	10		9
Gancao	Radix Et RhizomaGlycyrrhizae	11		10
Chaihu	Radix Bupleuri	12		11
Ezhu	RhizomaCurcumae		6	
Sanleng	RhizomaSparganii		8	
Zexie	RhizomaAlismatis			12
Daxueteng	Caulis Sargentodoxae		9	
Xiangfu	RhizomaCyperi			
Chuanshanjia	SquamaManis		10	
Haizao	Sargassum		11	

C. Merging the CHMs both in plain text format and formula

Since we already have the frequency list of herbs from either plain text format or formula, we can form a new table by putting them together. The merged results can be seen in Table 1. From the new order, we can find , in the top 12 merged herbs, 11 are from formula and 7 are from plain text format, and totally 8 herbs existing in both plain format and formula.

D . Network establishment

1) Network of CHMs: Based on the table of co-existed pairs of CHM, and in line with the principle of permutation and combination, we calculated the frequencies of herb pairs which existed in the whole dataset. Then sort them on descendent order of frequency, and this sorted list is the original data for network of CHMs filtered out from plain text format, and the same, we do this to the co-existed pairs in formula, and then accumulate the frequencies of pairs in both formats. Finally, we got merged co-existed pairs of CHMs.

$$\text{Herb pairs}_{(\text{merged})} = \text{Herb pairs}_{(\text{plain})} + \text{Herb pairs}_{(\text{formula})}$$

2) Network of symptom and pattern: As we calculated the list of CHMs, we also calculate the list of TCM patterns and symptoms associated with paper ID. Then, based on the principle of co-existing mentioned above, we build up the network of TCM pattern and symptom on OC. The pattern network and the symptom network are presented in Fig. 2.

III. Results

A . CHMs most frequently applied in curing OC:

Table I clearly indicates that the top five herbs we mined out from plain text format are Fuling, Guizhi, Dangdui, Baishao, Dahuang, and the top five herbs in formula are Fuling, Guizhi, Danpi, Taoren and Chishao. Finally, the top herbs after merging are Fuling, Guizhi, Taoren, Danpi, Chishao. They are all functioning in smoothing qi and removing blood stasis as well as alleviating water retention so as to resolving hard lumps.

B . Network construction:

1) Symptom network: In figure.2, the upper part shows the main symptoms of OC are "lower abdominal pain", "lump in the lower part of abdomen", "irregular menstruation", and "leukorrhagia", "dyspareunia". All these symptoms can be explained by pattern of qi stagnation and blood stasis, so we paint an arrow to reveal the relationship between them.
2) Pattern network:There is a distinguished pattern in the middle part of Figure.2; it is "Qi stagnation and blood stasis", which has extensive connection with other patterns.
3) Network of CHMs: In this part, the major herbs are marked by dark circles. The three circles are respectively formed three different formulas. Among them, Fuling, Guizhi, Danpi, Taoren and Baishao constituted "Guizhifulingwan", Chuanxiong, Danggui, Fuling, Baizhu and Chishao made up "Dangguishaoyaosan", also Dangshen, Fuling, Baizhu and Gancao composed "sijunzitang". Beside these mentioned herbs, there are four herbs not included in these three classical formulas. They are "Sanleng", "E'zhu", "Daxueteng" and "Haizao". These four herbs have common effect on strongly activating blood flowing and solving the lumps and they are also the popular herbs in OC curing.

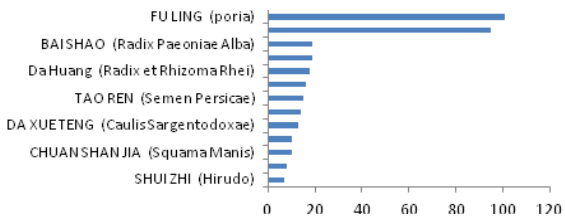

Fig 1. Top 13 herbs in OC in plain text format

IV . Discussion

Our study not only mined out the frequently used herbs in OC but also established the network of CHMs, symptoms and patterns for OC and marked the relationship among three of them. All those manipulations are supported by the technology of data mining, through which the specific rules in OC treatment can be figured out. The more details are as follows:

1) Text mining is a practicable way in excavating the co-existed pairs of CHMs and the merged results would be a more reasonable outcome for clinical practice. By merging the order of CHMs from plain text format and formula, we can get a new order that is more approaching the truth and highly related with written literatures. Formula is a treasure house, and it has been carried out in thousand years of practice. The characteristic of the formula is variety. Physicians always change the formula slightly to fit for the specific condition of the patients, so decomposing the formula and reordering the herbs would be a better way to help understanding the medication regularity.
2) Through slicing all relevant literatures we can obtain the major principles of OC treatment as well as the most frequently used herbs. From the network of pattern we can know that qi stagnation and blood stasis is the major pathology for OC and the etiology can be either phlegm, dampness and intertwined phlegm and stasis or deficient spleen qi. Both excess evil and deficient genuine qi can lead to qi stagnation and blood stasis [10]. Correspondingly, the frequently used formula are “Guizhifulingwan” or “Dangguishaoyaosan”, or “Sijunzitang”, which specialized at activating qi and blood, tonifying spleen qi and driving away lumps. The frequently presented herbs in CHMs network can constitute all above-mentioned formula also. In addition, scattered herbs in lower part of fig.2 are “Sanleng”, “E’zhu” “Daxueteng” and “Haizao”. They have strong power in smoothing qi and driving blood and are widely applied in gynecological diseases [11], although they are not grouped in to the formula. Putting all those information together, we can easily detect the logical relationship among the three networks. The correlation among three networks also agrees with the clinical practice in a high degree. Therefore, text mining will undoubtedly become a skillful method, which can figure out the treatment principle of various diseases and establish a basement for future research.

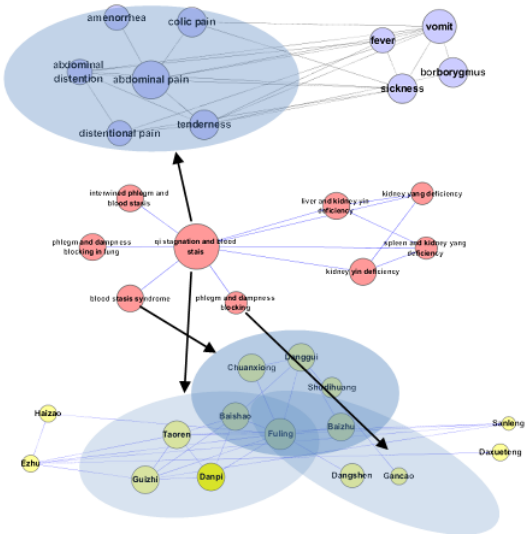

Fig2: Network relationship among herbs, symptoms and patterns for OC

Reference

[1] Robert T. Greenlee, Bruce Kessel, Prevalence, incidence and natural history of simple ovarian cysts among women over age 55 in a large cancer screening trial Am J Obstet Gynecol.vol. 202, no. 4, pp.373, 2010

[2] F Parazzini, S Moroni, E Negri, C La Vecchia, R Mezzopane, and P G Crosignani. Risk factors for seromucinous benign ovarian cysts in northern Italy.J Epidemiol Community Health. vol. 51, no. 4, pp. 449–452, 1997.

[3] Kim L. Thornton, Alan H. Decherney , Laparoscopic Management of ovarian cysts: An Endocrinologist View , The YALE Journal of Biology and Medicine. vol. 64, pp. 599-606,1991.

[4] GuangZheng, Miao Jiang, Xiaojuan He1, Jing Zhao, HongtaoGuo,Gao Chen, QinglinZha, and Aiping Lu. Discrete derivative: a dataslicing algorithm for exploration of sharing biological networks between rheumatoid arthritis and coronary heartdisease. BioDataMining.vol.4, pp.18, 2011

[5] Andreas Hotho, Andreas Nurnberger, and Gerhard Paaß, A Brief Surveyof Text Mining, LDV Forum - GLDV Journal for Computational Linguisticsand Language Technology, vol.20, no.1, pp. 19 - 62, 2005.

[6] D. J. Hand, HeikkiMannila, Padhraic Smyth, Principles of data mining,ISBN 026208290X, 9780262082907, MIT Press, 2001.

[7] George Tzanis, Christos Berberidis, Ioannis P. Vlahavas, Biological DataMining, Encyclopedia of Database Technologies and Applications. 2005.

[8] Sam Schmidt, Peter Vuillermin, Bernard Jenner, YongliRen, Gang Li,Yi-Ping Phoebe Chen, Mining Medical Data: Bridging the KnowledgeDivide, Proceedings of eResearch Australasia, 2008.

[9] GuangZheng, Junping Zhan, HongtaoGuo, MiaoJiang, Cheng Lu, and Aiping Lu. Rule-based Text mining of traditional Chinese medicine patterns with Chinese Herbal Medicines and Symptoms on Cirrhosis. In International Symposium on Information Technology in Medicine and Education. 2012.

[10] FuchunCi, Li Zhang, Analysis of pattern, symptom and Chinese herbal Medicine of Traditional Chinese Medicine curing Ovarian Cyst.vol. 24, no. 8, pp.20-23, 2011.

[11] Jinrong Fu, CongQi, The current situation of Traditional Chinese Medicine curing Ovarian Cyst, Journal of Traditional Chinese Medicine literatures. vol.3 pp.56-57, 2007.