FUZZY ANALYSIS OF THE CANCER RISK FACTOR*

Gabriela Dudek†, Anna Strzelewicz, Monika Krasowska
Aleksandra Rybak, Roman Turczyn

Department of Physical Chemistry and Technology of Polymers
Silesian University of Technology
Ks. M. Strzody 9, 44-100 Gliwice, Poland

(Received April 23, 2012)

A system which allows to predict type of cancer on the basis of the largest risk factors for particular patient, was created using fuzzy set theory. Lung, colon, breast, colorectal, stomach, cervical and prostate cancer were considered. The Mamdani model, implemented in the Fuzzy Logic Toolbox in Matlab, was used for data analysis. As inputs to the system genetic, biological (race, age, sex) and behavioral (overweight, alcohol consumption, tobacco smoke) risk factors were taken. The output was "the kind of cancer". Obtained results show that fuzzy logic can be an effective tool in dealing with this kind of medical problem.

DOI:10.5506/APhysPolB.43.947
PACS numbers: 87.15.Aa, 87.10.+e

1. Introduction

Nowadays, cancer is a major public health problem. Currently, one in four deaths is due to the cancer [1, 2]. Doctors often cannot explain why one person develops cancer and another does not. But research shows that certain risk factors increase the chance that a person will develop cancer [3]. A risk factor is a variable associated with an increased risk of disease. Factors that increase cancer risk can be external and internal. External factors include personal lifestyle, choices or substances (chemicals and asbestos) present in the environment, that are known to cause a cancer. Large group of people have also internal risk factors for cancer such as a genetic predisposition or those that develop during aging process. The most common risk factors for cancer are: growing older, tobacco, sunlight, ionizing radiation,

* Presented at the XXIV Marian Smoluchowski Symposium on Statistical Physics, "Insights into Stochastic Nonequilibrium", Zakopane, Poland, September 17–22, 2011.

† gmdudek@.polsl.pl

certain chemicals and other substances, some viruses and bacteria, certain hormones, family history of cancer, alcohol, poor diet, lack of physical activity or overweight [4,5]. Over time, several factors may act together to cause normal cells to become cancerous. Better estimation of cancer risk probability requires more intensive clinical services and research. By means of the fuzzy set theory this risk factors of cancer can be chosen very quickly. In this paper, we tried to use fuzzy set theory to create programme which can allow to predict the risk factors for different kind of cancer. Fuzzy sets theory [6,7,8] derives from the fact that almost all natural classes and concepts are fuzzy rather than crisp in nature. In traditional rule-based approaches, knowledge is encoded in the form of antecedent consequent structure. When new data are encountered, it is matched to the antecedents clause of each rule, and those rules, where antecedents match the data exactly are fired, establishing the consequent clauses. This process continues until the desired conclusion is reached, or no new rule can be fired. The proposed fuzzy logic approach is effective and become often used in many scientific areas [9, 10, 11, 12, 13]. It has been used in applications that are amenable to conventional control algorithms on the basis of mathematical models of the system being controlled, such as automatic control, data classification, decision analysis, expert system, and computer vision. Due to the inventor of fuzzy logic, medical diagnosis would be the most likely used domain of his theory *i.e.* in areas, where precise mathematical description of the control process is impossible and thus it is especially suited to support medical decision making. In this perspective, we described how fuzzy logic works, by illustrating its application in cancer's risk factors analysis. To get global overview on the problem we considered different types of cancer. Furthermore, we took data from different sources to observe how the risk factor changes with continent. The data for our model came from:

(a) *Global Cancer Statistics, 2002, A Cancer Journal for Clinicians* [14].

(b) *Alcohol Consumption and Lung Cancer, Cancer Epidemiology, Biomarkers and Prevention* [15].

(c) *Race/Ethnicity and Multiple Cancer Risk Factors among Individuals Seeking Smoking Cessation Treatment, Cancer Epidemiology, Biomarkers and Prevention* [16].

(d) *Hereditary Cancer Information*, `http://www.disabled-world.com` [17].

(e) *Alcohol Consumption and Risk of Colon Cancer: Evidence from the National Health and Nutrition Examination Survey I Epidemiologic Follow-up Study, Nutrition and Cancer* [18].

For every kind of cancer we looked for the information how risk factor has changed with sex, continent, age, heredity, smoke and alcohol. Some types of cancer concern only males (like prostate cancer) and some of them concern only females (breast, cervix). The results of our research are shown in Table I.

It seems interesting how the risk factor changes with the continent. Patients who live in Europe have the largest risk of breast, colorectal and lung cancer [19, 20]. Americans have the largest health hazard for prostate cancer [21, 22] but Asians — for stomach cancer [23, 24]. People who live in Africa have the largest risk of cervix cancer [14].

The primary risk factor for the disease with about 77 percent of all cancers is age. The association between age and an increased risk of cancer is not well understood. Certain changes in the cells could cause the beginning of cancer disease, but scientists cannot explain the reasons of this process. The influence of age on risk factor is different for various kinds of cancer. For example, males who are 50–90 years old have the largest risk of prostate cancer [22].

Many cancer diseases take place within the pale of the same family and the immediate relatives (siblings, parents, and children) of patients with cancers often have an increased risk of cancer [25, 26, 27, 28]. Breast cancer is an example in which immediate relatives may have a higher risk. Sometimes the increased risk of the cancer may be due to a genetic mutation. BRCA1 and BRCA2 are examples of genetic mutation that causes a rise of risk of breast cancer. Women who have this mutation have about 60% greater probability to develop breast cancer. A similar situation is in the case of colon cancer; there are specific identifiable genetic mutations that give rise to very high risks of developing colon cancer. Mutation in the APC gene is a noted example.

The fact that cigarettes contain cancer causing chemicals is nothing new [16]. For years, cancer experts have been advising the public to get out of the habit and decrease their cancer risk. The connection between alcohol consumption and cancer risk, however, is not so well known. When alcohol consumption is combined with smoking, the incidence of cancer is even higher. Probably, cells dehydrated by alcohol are more sensitive to the effect of cigarette's chemicals, causing cancer [18].

Gathering information about sex, continent, age, heredity, smoke and alcohol allows creating the reliable programme for predicting the highest probability of cancer kindthreatening patients. This evaluation is necessary to decide which kind of medical tests should be done for an individual patient.

TABLE I

Cancer statistics.

The kind of cancer	Sex [%]		Continent [%]				Age	Heredity [%]	Smoke	Alcohol
	Males	Females	Africa	America	Asia	Europe				
Breast	0,00	17,90	13,73	19,53	15,05	31,02	20–80	80	no influence	1.5 times more
Cervix	0,00	5,72	35,41	15,76	15,43	10,41	35–55	20–40	influence	2,0 times more
Colorectal	6,15	5,34	7,38	16,98	20,66	33,86	40–79	20	influence	influence
Lung	3,81	1,72	7,12	18,37	24,47	32,79	51–84	50	the largest influence	1,9 times more
Prostate	9,60	0,00	15,55	29,56	5,46	25,72	50–90	10	no influence	influence
Stomach	3,86	2,12	12,83	14,52	37,74	21,24	40–70	50	influence	no influence

2. Mamdani fuzzy interference system

Our programme was created on the basis of Mamdani fuzzy model. A fuzzy model expresses a complex system in the form of fuzzy implications. Mamdani model [29] can be built by using these implications (linguistic relationships) and observed data.

In the Mamdani fuzzy model, we define X as input (regression) matrix and y as output vector in the following form

$$X = [x_1, \ldots, x_n]^T = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix}, \tag{1}$$

$$y = [y_1, \ldots, y_n] \,. \tag{2}$$

General form of linguistic fuzzy *if–then* rules is given as follows

$$R_i : \text{if } x \text{ is } A_i \text{ then } y \text{ is } B_i \,, \qquad i = 1, 2, \ldots, K \,, \tag{3}$$

where R_i is the rule number, A_i and B_i are the fuzzy sets, x is the antecedent variable representing the input in the fuzzy system, and y is the consequent variable related to the output of the fuzzy system.

The general algorithm of Mamdani fuzzy inference system can be presented in the following three steps (Fig. 1):

(a) Computation of the degree of fulfillment B_i of the antecedent for each rule i:

$$\beta_i = \mu_{Ai1}(x_1) \wedge \mu_{Ai2}(x_2) \,, \qquad 1 \leq i \leq K \,; \tag{4}$$

(b) Derivation (for each rule) the output fuzzy set B_i' using the minimum t-norm:

$$\mu_{B_i'}(y) = \beta_i \wedge \mu_{B_i}(y) \,; \tag{5}$$

(c) Aggregation of the output fuzzy sets by taking the maximum (union):

$$\mu_{B'}(y) = \max_{i=1,2,\ldots,K} \left(\mu_{B_i'}(y) \right) \,. \tag{6}$$

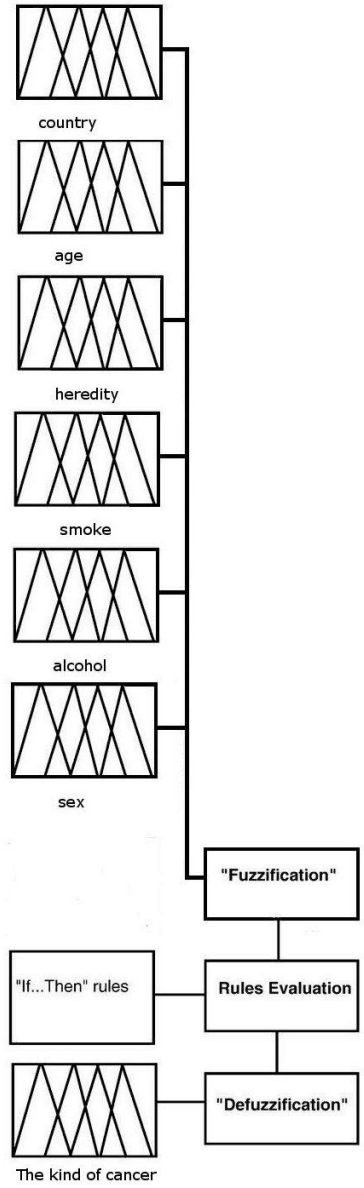

Fig. 1. General description of fuzzy expert to analysis the risk factor of cancer.

The *max* operator is the most common implementation of the rule aggregation operation. According to this procedure, the overall fuzzy output is calculated from the set of individual outputs taking the maximum truth value, where one or more terms overlap [30].

To obtain estimated (crisp) value $y*$, the output subset must be defuzzified. There are many different methods of defuzzification, described in [31], such as: Center of Gravity(COG), Centroid Average (CA), Maximum Center Average (MCA), Mean of Maximum (MOM), Smallest of Maximum (SOM), Largest of Maximum (LOM).

In our case we used the Mean of Maximum (*Middle of Maxima*) method to defuzzify the output, excepts that the locations of the maximum membership can be non-unique. Computing the crisp output of the system as the maximum value, finally, we obtain fuzzy set. When there is more than one output value, the crisp values are averaged [32].

3. Application of the method

Collected data consist of six inputs. The six attributes detailed in Table II are graded on an interval scale from a normal state of 0 to 1.

TABLE II

Cancers data: description of inputs and output.

Inputs No.	Inputs description	Membership functions description	Minimum	Maximum	Mean
1	Continent	Africa	$-0,33$	0,33	0,00
		Asia	$0,00$	0,67	0,33
		America	$0,33$	1,00	0,67
		Europe	$0,67$	1,33	1,00
2	Age	0–20	$-0,25$	0,25	0,00
		20–40	$0,00$	0,50	0,25
		40–60	$0,25$	0,75	0,50
		60–80	$0,50$	1,00	0,75
		80–100	$0,75$	1,25	1,00
3	Heredity	small influence	$-0,50$	0,50	0,00
		middle influence	$0,00$	1,00	0,50
		big influence	0,50	1,50	1,00
4	Smoke	no influence	$-0,50$	0,50	0,00
		influence	$0,00$	1,00	0,50
		the largest influence	$0,50$	1,50	1,00
5	Alcohol	no influence	$-0,50$	0,50	0,00
		influence	$0,00$	1,00	0,50
		the largest influence	$0,50$	1,50	1,00
6	Sex	males	$-0,01$	0,01	0,00
		females	$0,99$	1,10	1,00

The first input is continent and it contains four membership functions: Africa, Asia, America and Europe.

We have divided the second input — age — into five membership functions: 0–20 years, 20–40 years, 40–60 years, 60–80 years and 80–100 years.

Heredity is the third input with: small influence, middle influence and big influence membership functions.

The other two inputs are smoke and alcohol for which the membership functions are: no influence, influence and the largest influence.

The last input is sex, of males and females as membership functions.

Total number of rules (R_i) is 1286, some chosen are shown in Table III.

TABLE III

Fuzzy rules.

Rule No.	Continent	Age [year]	Heredity	Smoke	Alcohol	Sex	The kind of cancer
1	Africa	20–40	big influence	no influence	the largest influence	female	breast
...							
22	America	20–40	middle influence	influence	the largest influence	female	cervix
...							
56	Europe	40–60	small influence	influence	influence	male	colorectal
...							
123	Asia	60–80	middle influence	the largest influence	the largest influence	female	lung
...							

The method of interpretation of chosen rules (from Table III) is given below.

Rule 1: IF the continent is Africa, the patient is a female, between 20–40 years old, with big influence of heredity, who does not smoke and drinks a lot of alcohol: THEN patient is endangered with the largest risk of getting the breast cancer.

Rule 22: IF the continent is America, the patient is female, between 20–40 years old, with middle influence of heredity, who smokes and drinks a lot of alcohol: THEN patient is endangered with the largest risk of getting the cervix cancer.

Rule 56: IF the continent is Europe, the patient is male, between 40–60 years old, with small influence of heredity, who smokes and drinks an alcohol: THEN patient is endangered with the largest risk of getting the colorectal cancer.

Rule 123: IF the continent is Asia, the patient is female, between 60–80 years old, with middle influence of heredity, who smokes and drinks a lot of alcohol: THEN patient is endangered with the largest risk of getting the lung cancer.

4. Results

The collection of well-distributed, sufficient, and accurately measured input data is the basic requirement to obtain an accurate model. The data sets, consisting of 6 inputs, are summarized in Table I. These data were implemented in MATLAB as Mamdani fuzzy interference system. In order to select the best model among the tested fuzzy variable combinations, we have used Gaussian membership function for the inputs and triangular membership function for the output. These membership functions are simple and precise in determining the value of the input parameters. As the defuzzification method we have used Mean of Maximum (MOM).

The membership function and set of rules are fed to the system to determine the output. Each rule in the system is very important and critical for generation of the predictions in numeric form. The graphs of the membership function for "sex" and "age" are shown in Fig. 2 and 3, respectively. Taking into account different combinations of inputs and 1286 rules we could check reliably and satisfactorily for what kind of cancer we have the largest risk.

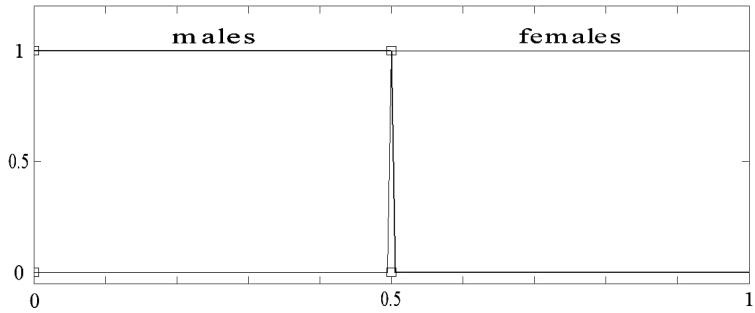

Fig. 2. Sex MF for fuzzy model.

The surface models with two significant parameters could be used to show two ways of interactions and relationships towards the desired response. The examples of such surface analysis for interactions between continent and sex,

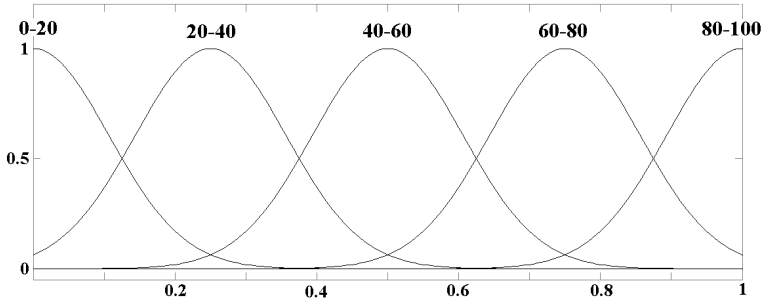

Fig. 3. Age MF for fuzzy model.

and between continent and age are shown in Fig. 4 and 5. The proportional and non-proportional relationships of the inputs towards the desired output are clearly visible in these figures. They could be used to study interaction effects in achieving high quality programme to prediction risk factor for some kind of cancer.

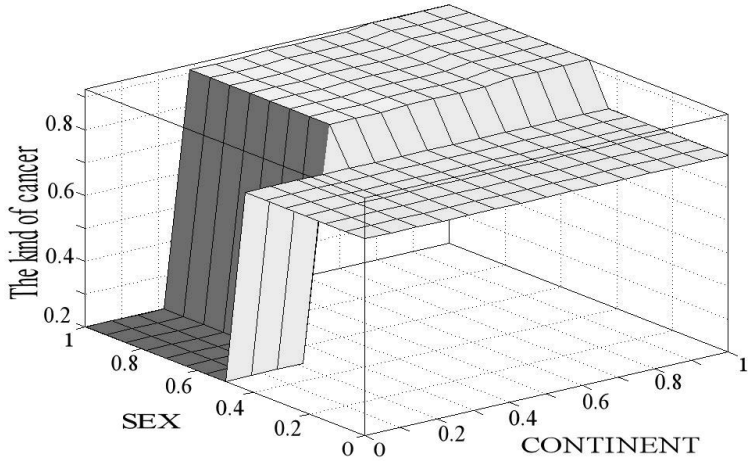

Fig. 4. Surface analysis between continent and sex.

As can be seen in figure 4 the largest risk of colorectal cancer endangers males irrespectively of their place of residence. In the case of females, the biggest risk of cervix cancer is in Africa. In other continents (America, Asia, Europe) we can observe the largest risk for breast cancer.

In figure 5 we can see various relationships between age, continent and the kind of cancer.

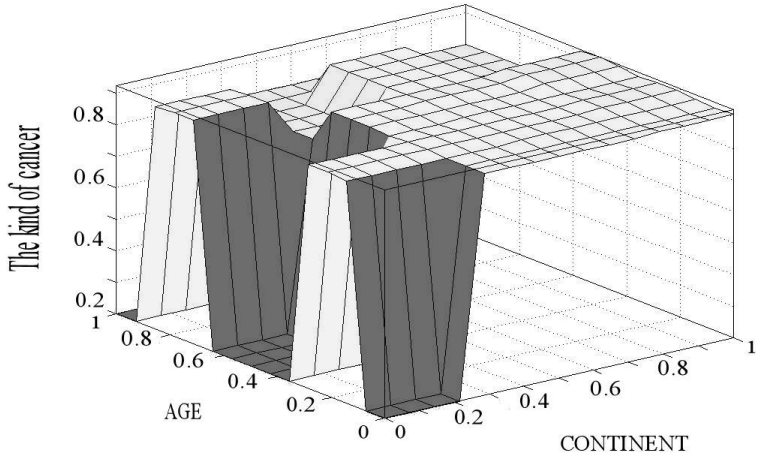

Fig. 5. Surface analysis between continent and age.

Females between 0–15, 35–70 and 90–100 years old are the most endangered with the cervix cancer in Africa. In the case of breast cancer, the biggest risk exists for females between 15–35 and 70–90 years old. For the Asian continent the largest risk exists for breast (0–65 years old) and colorectal (65–100 years old) cancer. We can observe the same situation for the Europe. The greatest risk of breast cancer exists for the American people between 0–90 years old, while the people above 90 years old are endangered by colorectal cancer.

TABLE IV

Interaction between inputs and output for some data.

Continent	Age	Heredity	Smoke	Alcohol	Sex	The kind of cancer
1,0	0,5	0,9	0,1	1,0	1,0	Breast
0,5	0,7	1,0	0,1	1,0	1,0	Breast
0,3	0,5	0,1	0,5	0,5	0,0	Colorectal
0,7	0,3	0,0	0,1	0,1	1,0	Colorectal
0,7	1,0	0,0	0,0	0,5	0,0	Prostate
1,0	0,4	0,0	1,0	1,0	0,0	Stomach and prostate
0,0	0,6	0,5	0,1	0,0	1,0	Stomach
0,0	0,7	0,5	0,1	0,1	1,0	Stomach
0,0	0,4	0,0	0,1	1,0	1,0	Cervix
1,0	0,4	0,2	0,5	1,0	1,0	Cervix
0,5	0,7	0,5	1,0	1,0	1,0	Lung
0,3	0,7	0,5	1,0	1,0	0,0	Lung

The set of some inputs in a numeric form with their proper response is shown in Table IV. In the example below, we describe haw to interpret the results from the table. Suppose that the patient is a female, who lives in Europe and is 50 years old, does not smoke, but drinks alcohol and her heredity has a big influence. We can see that she has the largest risk of getting the breast cancer (see Table IV).

5. Concluding remarks

In this paper, we created the programme for estimation of risk factors for breast, cervix, colorectal, lung, stomach and prostate cancer. The programme is based on Mamdani fuzzy system, which is often used to support medical decision making. The proposed fuzzy logic approach has the potential and opens a new and promising direction for effective and early treating of different kinds of cancer. Our programme allows to predict a type of cancer for which patient has the largest risk factor. The system contains six inputs: continent, age, heredity, smoke, alcohol and sex. Basing on the given inputs and general algorithm of Mamdani fuzzy system we obtain 1286 rules. The output is "the kind of cancer" with the biggest risk factor.

The most significant parameters that affect the risk factor are "sex" and "continent". In the case of females, the largest risk factor is for breast cancer in the most continents. Only in Africa the biggest risk factor is for cervix cancer. Whereas in the case of males irrespectively of their place of residence the largest risk factor is for colorectal cancer.

Our programme could find wide application in medicine diagnostic. Our investigations have shown that fuzzy logic method gives more effective and useful classifications results. We are going to work on a wider source of data which will allow to develop the programme.

The authors would like to thank Prof. Zbigniew J. Grzywna for substantial help and the Silesian University of Technology for providing financial support under the project No. GG 103/CZ2/2011.

REFERENCES

[1] A. Jemal *et al.*, *CA Cancer J Clin.* **58**, 71 (2008).

[2] National Center for Health Statistics, Division of Vital Statistics, Centers for Disease Control and Prevention. Available at: `http://www.cdc.gov/nchs/nvss.htm` (2007).

[3] D. Coggon, C.N. Martyn, *The Lancet* **365**, 1434 (2005).

[4] J. Cielecka-Piontek, A. Styszynski, K. Wieczorowska-Tobis, *New Medicine* **1**, 2 (2004).

[5] M. Ezzati *et al.*, *The Lancet* **360**, 1347 (2002).
[6] L. Zadeh, *Int. J. Man-Machine Studies* **8**, 249 (1976).
[7] L. Zadeh, *Information Control* **8**, 338 (1965).
[8] L. Zadeh, *IEEE Trans. on Knowledge and Data Engineering* **1**, 89 (1989).
[9] G. Dudek, Z.J. Grzywna, M.L. Willcox, *Biosystems* **94**, 285 (2008).
[10] H. Yao, L.-Y. Shen, J. Hao, C.M. Yam, *Management of Environmental Quality: An International Journal* **18**, 442 (2007).
[11] T. Morato, W.W.L. Cheung, T.J. Pitcher, *Journal of Fish Biology* **68**, 209 (2006).
[12] S. Lekkas, L. Mikhailov, *Artif. Intell. Med.* **50**, 117 (2010).
[13] T.P. Exarchos *et al.*, *Artif. Intell. Med.* **40**, 187 (2007).
[14] D. Parkin, F. Bray, J. Ferlay, P. Pisani, *CA: A Cancer Journal for Clinicians* **55**, 74 (2005).
[15] E.V. Bandera, J.L. Freudenheim, J.E. Vena, *Cancer Epidemiol. Biomarkers Prev.* **10**, 813 (2001).
[16] D.E. Kendzor, T.J. Costello, Y. Li, *Cancer Epidemiol. Biomarkers Prev.* **17**, 2937 (2008).
[17] Hereditary Cancer Information Available at: `http://www.disabled-world.com`
[18] L.J. Su, L. Arab, *Nutr. Cancer* **50**, 111 (2004).
[19] M.M. Center, A. Jemal, R.A. Smith, E. Ward, *CA: A Cancer Journal For Clinicians* **59**, 366 (2009).
[20] F. Hrubá *et al.*, *Central European Journal Of Public Health* **17**, 115 (2009).
[21] S. Loeb, E.M. Schaeffer, *Primary Care* **36**, 603 (2009).
[22] J.L. Watters *et al.*, *Cancer Epidemiol. Biomarkers Prev.* **18**, 2427 (2009).
[23] H. Brenner, D. Rothenbacher, V. Arndt, *Methods In Molecular Biology* **472**, 467 (2009).
[24] B. Sumathi, S. Ramalingam, U. Navaneethan, V. Jayanthi, *Singapore Medical Journal* **50**, 147 (2009).
[25] B.S. Hulka, P.G. Moorman, *Maturitas* **104**, 203 (2008).
[26] H. Jefferies, *Nursing Times* **104**, 26 (2008).
[27] S.S. Rao, M. Singh, M. Parkar, R. Sugumaran, *American Family Physician* **78**, 583 (2008).
[28] P. Parsa, B. Parsa, *Asian Pacific Journal Of Cancer Prevention* **10**, 545 (2002).
[29] E.H. Mamdani, *Fuzzy Sets Syst.* **26**, 1182 (1977).
[30] D.T. Pham, M. Castellani, *Proc. Instn. Mech. Engrs. Part. C.: J. Mechanical Engineering Science* **216**, 747 (2002).
[31] J.J. Saade, H.B. Diab, *Man, and Cybernetics. Part B: Cybernetics, IEEE Transactions* **30**, 223 (2002).
[32] I. Rojas, O. Valenzuela, M. Anguita, A. Prieto, *Internat. J. Approx. Reason* **19**, 367 (1998).