Analyzing a Substructure of Phenoxyherbicide Production Workers Cohort by the Pattern Recognition

Zarema Amirova

Environmental Research Center of the Republic of Bashkortostan, 147, October Av., 450075

Introduction

Several criteria are used to form cohorts of dioxin impact: probable occupational exposure of phenoxyherbicide production workers (2,4,5-T, PCP of more than 10 tons per year) or chloracne case [1]. These criteria are used when there is no analytical data on PCDD/Fs levels in biological tissues of potentially exposed workers. Most reliable subcohort is formed out of experimental data array by the TEQ value of blood or fat sample. However in this case the criteria are somewhat fuzzy because the application of the standard parameter TEQ does not permit to compare the concentration profiles of samples, and TEF sharply decreases statistical weight of polychlorinated isomers. The notion of norm in dioxin pollution is not defined. Pollution is considered to be high if it by 2-5 times exceeds the background level of control [2]. This approach can be used in case of homogeneous pollution. However when analyzing the structure of a joint cohort of Ufa workers we faced a more complicated situation. Workers in Ufa were exposed to various pollutants. Chloracne was diagnosed to 128 persons for a short time involved in 2,4,5-T production [3]. Besides this group there was a group of exposed workers for a long time involved in 2,4,5-TCP, 2,4-D, MCPA and other chlororganic productions with no signs of occupational disease. The majority of these workers were exposed to the pollution produced by equipment, products and the plant territory. This resulted in the fact that the isomer specter for different groups differs considerably: for 2,4,5-T workers the main pollutant is 2,3,7,8-TCDD, for 2,4-D and TCF workers – hexa- and octachlorinated isomers. The difference was up to 50-150 times depending on the isomer though TEQ value deviation did not exceed 1.5 times. Now for the majority of these groups' representatives the current PCDD/Fs content in blood is known what permits to apply methods of multidimensional statistics to formalization of the criteria for subcohort formation.

Objects and methods of data analysis

In 1996-1998 PCDD/Fs content was determined in blood of 75 workers and 44 unexposed donors (N). Table gives mean data on PCDD/Fs content in this joint cohort consisting of workers who produced 2,4,5-T (T, n=41), 2,4,5-TCP (O, n=4), 2,4-D (D, n=24), chlorbenzene and MCPA (R, n=6) and 4 synthesis chemists (S) who carried out 2,3,7,8-TCDD pilot run of production.

The methods of cluster analysis and recognition of fuzzy sets estimated the observed differences of PCDD/Fs isomer specter in blood samples of exposed and unexposed donors. The initial set of experimental data was $\{X_{ij}\}$, $i=1 \div n$, $j=1 \div m$, where n is a number of characters (concentration of 17 PCDD/Fs isomers), m is a number of objects (blood samples of donors) consisting of subsets $\{X_{ij}\} = \{T_{ij}\} \cup \{O_{ij}\} \cup \{D_{ij}\} \cup \{R_{ij}\} \cup \{S_{ij}\} \cup \{N_{ij}\}$. It was necessary to find interrelation between these subsets.

TableMain parameters of risk groups and a control group of donors in Ufa, 1996/98

PCDD/Fs	T(n=41)	O(n=4)	D(n=24)	R(n=6)	S(n=4)	N(n=44)
2378-TCDD	103.9	318.6	98.1	53.9	647.4	21.8
12378-PnCDD	39.6	265.4	164.4	89.6	27.4	13.6
123478-HxCDD	16.6	67.0	70.9	14.7	8.3	6.6
123678-HxCDD	38.9	151.6	190.8	42.7	20.3	10.7
123789-HxCDD	19.9	53.5	46.0	29.3	12.7	5.6
1234678-HpCDD	52.8	39.1	140.8	57.4	24.4	17.4
OCDD	209.0	216.8	138.9	526.2	106.9	92.3
2378-TCDF	11.3	7.6	11.8	9.2	7.5	6.4
12378-PnCDF	10.8	17.4	14.1	8.9	13.4	10.2
23478-PnCDF	31.2	25.6	54.6	26.4	21.7	18.9
123478-HxCDF	24.9	50.3	45.2	18.8	11.7	13.8
123678-HxCDF	16.9	18.1	32.5	14.2	5.0	7.5
123789-HxCDF	7.7	6.9	9.0	3.4	4.0	5.7
234678-HxCDF	12.2	7.4	12.6	5.7	5.1	5.9
1234678-HpCDF	59.01	34.0	141.5	88.6	18.4	16.0
1234789-HpCDF	9.9	8.7	10.9	3.3	7.9	7.3
OCDF	60.6	73.2	58.3	20.3	43.7	26.2
TEQ, pg/g lipids	156.8	490.3	243.3	128.3	672.4	43.5

In the procedure of cluster analysis the Chebyshev distance between the elements was used as a similarity measure permitting to differentiate between the elements of the set if they significantly differ at least by one characteristic. The final number of clusters was determined by a change in the similarity measure d_{min} while passing from one level of hierarchy to another. To select the most typical representatives of subgroups we analyzed the substructure of the joint cohort of exposed donors by the method of fuzzy set recognition. The procedure of recognition implies a stage of a priori reference of a set of objects to some subsets, in the given case – to the groups selected by their occupation – the workers of 2,4,5-T, TCP, 2,4-D, MCPA, chlorbenzene, 2,3,7,8-TCDD production and also unexposed donors.

In the algorithm of fuzzy sets recognition a threshold logical element is used as a classifier, the belonging of the objects x_{ij} $(i=1\div n)$ to one of the pattern is determined by calculating the distance of the characteristics $(j=1\div m)$ in Euclid space from the object to the centers of alternative classes (A and B).

If $\rho_{xi,A'} < \rho_{xi,B'}$, then $x_i \in A$, otherwise, $x_i \in B$.

$$\rho_{xi,A'} = \left(\sum (x_{i,j}-A'_j)^2 \right)^{1/2}, \quad \rho_{xi,A'} = \left(\sum (x_{i,j}-B'_j)^2 \right)^{1/2};$$

$$A'_j = \sum_{j=1-n} x_{i,j}/m_{A,}, (i \in A); \quad B'_j = \sum_{j=1-n} x_{i,j}/m_{B,}, (i \in B); \qquad m_A + m_B = m.$$

Enforcing the notion of typicalness (closeness to the center of one's "own" class) by the notion of maximum remoteness from the center of the "anti-class" let's introduce a notion of "leader". The conditions of leadership are as follows: $L\ (\min \rho_{xi,A'}\ ;\ \max \rho_{xi,B'})$. The distance to the "leader" - $\rho_{xi,L}$ was used as a measure of objects' perspective. Recognition errors were estimated as a relation of mismatches of the received classification and the a priori set classification to the total number of objects of the analyzed classes.

Results and Discussion

As a result of the cluster analysis procedure in the variant of agglomeration hierarchic grouping of standard initial data (PCDD/Fs in the blood samples of exposed and unexposed donors) a true division into 3 clusters was received:

Cluster I	***Cluster II***	***Cluster III***
2,3,7,8-TCDD=181pg/g	2,3,7,8-TCDD=73.4pg/g	2,3,7,8-TCDD=181pg/g
TEQ=284 pg/g lipids	TEQ=135 pg/g lipids	TEQ=284 pg/g lipids
9n⊂T; 8n⊂D; 3n⊂O	13n⊂T; 2n⊂D; 1n⊂R; 9n⊂N	31n⊂N; 18n⊂T; 5n⊂R; 2n⊂D; 1n⊂O

As the dendrogram shows the class structure is complicated, 3 formed clusters have different levels of mean exposure but the selected clusters include elements from different groups. Only the subset of highly exposed donors (cluster I) does not overlap the subset N of unexposed donors. Next step is a cluster analysis of the subset of exposed workers of the joint cohort:

Cluster IV	***Cluster V***	***Cluster VI***
2,3,7,8-TCDD=192pg/g	2,3,7,8-TCDD=73.4pg/g	2,3,7,8-TCDD=181pg/g
TEQ=345 pg/g lipids	TEQ=135 pg/g lipids	TEQ=284 pg/g lipids
9n⊂T; 7n⊂D; 3n⊂O	13n⊂T; 2n⊂D; 1n⊂R; 9n⊂N	31n⊂N; 18n⊂T; 5n⊂R; 2n⊂D; 1n⊂O

The clustering permitted to single out 3 subcohorts with most typical isomer specters that are characterized PCDD/Fs levels by 3-4 and 9 times exceeding the background.

As it follows from the substructure analysis of the joint cohort it is impossible to single out any of the groups to which the samples are referred a priori. There is no reason to single out subcohorts with an a priori assumption of occupational exposure, at least now. This is referred to the subcohort 2,4,5-T in which chloracne was diagnosed. A better ground is the total level of experimentally confirmed exposure taking into account not only the TEQ sample but also the concentration "picture" of isomer composition of PCDD/Fs traces in blood. For consideration of this peculiarity we used data normalization: $x'_j=x_j/x_{imax}$ for $i=1 \div 17$ and $j \in T \cup R \cup D \cup S \cup N$. Representation of the structure of the analyzed set by a Venn diagram is given in Figure 1.

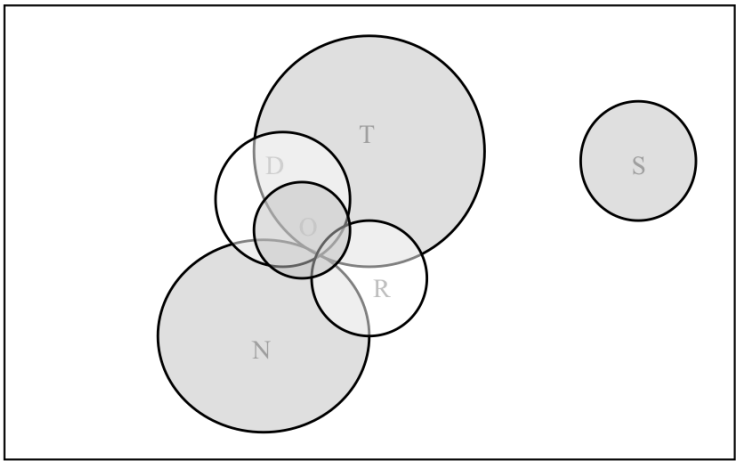

Figure 1. The structure of the set the results of PCDD/Fs determination in blood samples

The mean level of recognition of subsets T, O, R, D is 87%, but classes O and D can not be identified by the character of isomer specter (O∈D). Group S is an entirely isolated subset considerably different from the others.

Using the distance to the "leader" as a measure of objects "typicalness" we have found the dependence of TEQ sample on $\rho_{xi,L}$ describing experimental data of exposed donors (Figure 2). The dependence may be used for classification of new objects, for determination of "normal" and increased level, for selection of TEQ objects out of the uncertainty zone, for formalization of the criteria for forming subcohorts of highly exposed donors.

The assessment of remote consequences of PCDD/Fs exposure was carried out for the subcohort of highly exposed donors formed on the basis of the described approach [4].

References

1. WHO, IARC Monographs, 1996, 69, 53
2. Beck H, Escart K, Mathar W, Wittkovski R; Chemosphere, 1989, 18,507
3. Schecter A, Dioxins and Health, Plenim Press, NY, 1994, 466
4. Amirova Z. et.al. Presented in Dioxin-99

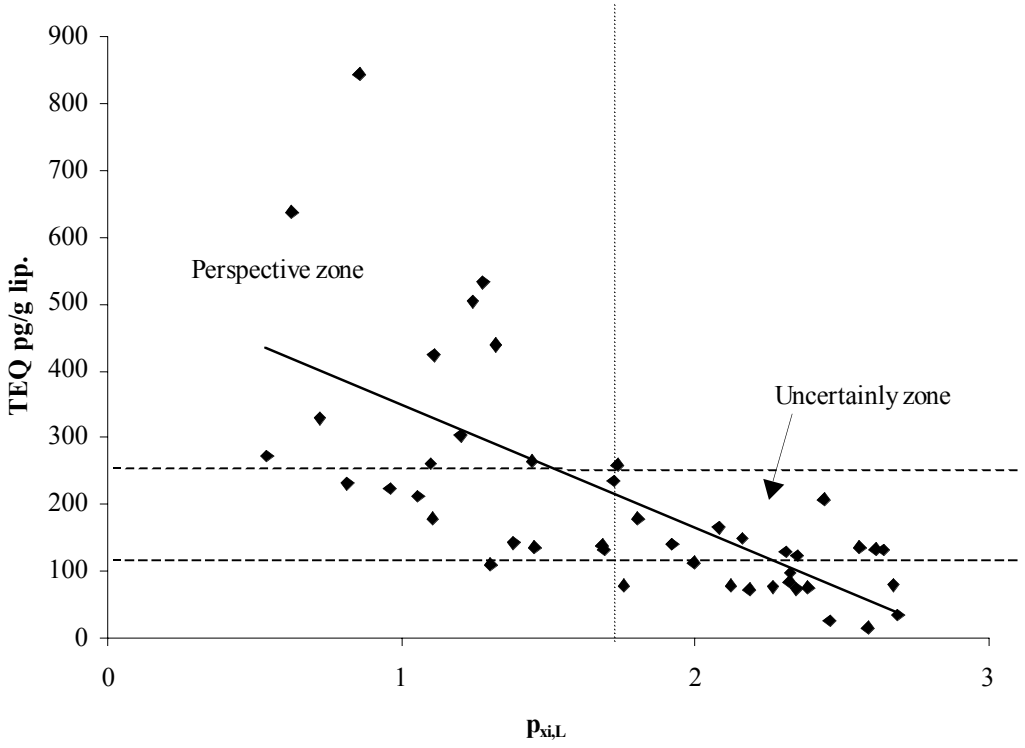

Figure 2. TEQ-distance from "leader"