Journal of Biomedical Informatics 44 (2011) 545–552

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Occupational exposome: A network-based approach for characterizing Occupational Health Problems

Laurie Faisandier [a,*], Vincent Bonneterre [a,b], Régis De Gaudemaris [a,b], Dominique J. Bicout [c,*]

[a] Laboratoire Environnement et Prédiction de la Santé des Populations, UMR CNRS 5525 Université Joseph Fourier, Domaine de la Merci, 38 706 La Tronche, France
[b] Service de Médecine et Santé au Travail, Centre Hospitalier Universitaire Grenoble, BP217, 38 043 Grenoble, France
[c] Biomathématiques et Epidémiologie, Laboratoire Environnement et Prédiction de la Santé des Populations-TIMC, UMR CNRS 5525 Université Joseph Fourier, VetAgro Sup, Campus Vétérinaire de Lyon, 1 avenue Bourgelat, 69280 Marcy l'Etoile, France

ARTICLE INFO

Article history:
Received 18 April 2010
Available online 27 February 2011

Keywords:
Occupational disease
Multi-exposure
Network

ABSTRACT

Surveillance of work-related diseases and associated exposures is a major issue of public health, in particular for identifying and preventing new threats for health. In the occupational health context, the French national occupational disease surveillance and prevention network (RNV3P) have constructed a growing database that records every year all Occupational Health Problems (OHPs) diagnosed by a network of physician specialists. The network aims to provide and develop an expertise on the disease–exposure relationships, and uses the RNV3P database for developing the surveillance of OHPs and for the detection of emerging associations between diseases and occupational exposures. In this paper, we have developed the theoretical framework of the occupational exposome, defined as a network of OHPs linked by similar occupational exposures, as a novel approach which allows to characterize and to analyze the disease–exposure associations reported in the RNV3P database in the form of a relational network. Next, the occupational exposome is structured in terms of occupational exposure groups which constitute informative sub-sets of hazards considered as the backbone tree spectrum of the occupational exposures potentially related to a disease.

To illustrate the wide possibilities of this method, the exposome approach is applied to the RNV3P database's sample of Non-Hodgkin Lymphomas (NHLs). As a result, we found that the NHL occupational exposome could be described in terms of 86 embedded exposure groups, defined as a set of OHPs sharing at least one component of the occupational multi-exposure. For example, "organic solvents and thinners" is the most represented hazards related to NHLs, but is also co-associated to "benzene", "ionizing radiations" or "agricultural products". From the knowledge stored in the database by physician experts, the occupational exposome constitutes a decisive step towards the evolving monitoring of multi-exposure associated to a given disease.

1. Introduction

Studies on health at work are designed primarily to identify occupational hazards, validate new atmospheric or biological indicators of exposure and effect, assess the impact of the occupational environment on health, gain information on the etiology of diseases and improve means of preventions. Surveillance of work-related diseases and exposures is a major issue of public health, in particular for identifying and preventing new threats for health. In such a context, surveillance conventionally involves in specific cohort and epidemiological follow-up of indicators (accidents and occupational diseases), often developed from surveillance networks which feed databases. An illustrative example of such a network is The Health and Occupational Reporting network (THOR) in the United Kingdom [1–3].

In France, the national occupational disease surveillance and prevention network (Réseau National de Vigilance et de Prévention des Pathologies Professionnelles, RNV3P) was set up in 2001 as a nation-wide network of experts from occupational health consultation centers in University Hospitals of metropolitan France that records in a structured and standardized RNV3P database all patient cases diagnosed with an Occupational Health Problem (OHP), i.e. patients with diseases potentially related to occupational exposures. The RNV3P network originated from the need of developing a common expertise on health problems by bringing together scientists and researchers working in multidisciplinary areas of health. Among others objectives, the network aims to provide and develop an expertise on the disease–exposure relationships, and uses the RNV3P database for developing the

* Corresponding authors. Fax: +33 4 76 76 89 10.
E-mail addresses: LFaisandier@chu-grenoble.fr (L. Faisandier), VBonneterre@chu-grenoble.fr (V. Bonneterre), RDegaudemaris@chu-grenoble.fr (R. De Gaudemaris), Dominique.Bicout@imag.fr, d.bicout@vetagro-sup.fr (D.J. Bicout).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.02.010

surveillance of OHPs and for the detection of emerging associations between diseases and occupational exposures. To address these issues, a systematic data-mining approach based on statistical tests of disproportionate as used in pharmacovigilance [4,5] was applied to RNV3P database to generate pre-alerts of potentially emergent disease–exposure couples. Although very sensitive, this method appears to be less specific and fails to simultaneous handle all components of the exposure associated with a disease.

Following these first studies on the RNV3P database, we began to develop the occupational exposome approach [6,7], as an alternative and complementary approach to pharmacovigilance methods involving disease–single exposure couples, that incorporates all dimensions of the composite occupational exposures in analyses of the RNV3P database. Our aim is to develop an approach allowing to investigate characteristics or traits which gather or separate OHPs as many individual factors and complex occupational situations, where exposures of diverse origins and variable intensity over time, are combined in effects on health. In this framework, we conceptualized the occupational exposome as a network of OHPs sharing components of the set of occupational exposures. Such an occupational exposome will make sense only when dealing with diseases associated with multi-exposures, i.e., associated with more than one hazard.

Similar conceptual approaches can be found in literature including the wording exposome which had been already introduced by Wild [8]. In contrast to the occupational exposome, the Wild's exposome represents the collection and succession of individual and environmental exposures encountered during an individual's lifetime. The author proposes reconstructing for a given individual, an exposure network in order to better understanding the role of each exposure and thus generate research hypotheses for the disease etiology. Although this requires a certain amount of information on traceability of exposures, this exposome view lays the groundwork for a coherent debate geared towards the development of relational networks for monitoring of exposures of various origins. Beyond the idea of collecting multi-exposures, Barabasi [9] has constructed a "diseasome" to illustrate, under a network, the environmental and social factors which might have a potential role in the origins of obesity; and Goh et al. [10] have explored a network of human diseases that implicates similar genetic mutations. In a similar way, Christakis et al. [11] have used the concept of relational networks that link individuals sharing social ties to monitor their weight over time. Assuming that the weight of an individual may be influenced by his or her surroundings, this approach has highlighted the possibility that social networking could be a factor in the spread of obesity. What all these analyses and various views have in common is the search of similar characteristics for understanding underlying mechanisms and identifying key factors in the onset and development of diseases.

Our main goal is to develop a framework allowing analyzing the RNV3P database in terms of an evolving and growing complex network. For this purpose, objectives of this paper are twofold: first, we outline the framework of the occupational exposome approach by providing definitions and showing how to construct exposomes, and second we show how to characterized and structured the topology of the constructed exposome by means of occupational exposure groups (sets of nodes or OHPs in a network sharing similar characteristics) which represent motifs carrying information on occupational exposures potentially related to the disease. To illustrate how this approach can be implemented, we considered the sample of Non-Hodgkin Lymphomas from the RNV3P database and, subsequently, we show the potential of this approach for monitoring and studying associations and relationships between diseases and occupational multi-exposures.

2. Material

Each diagnosed patient is recorded in the RNV3P database as an Occupational Health Problem (OHP) represented by a four codes item (Fig. 1) that includes a disease or pathology code (ICD-10 classification) plus a three-item exposure code holding information on the kind of industrial products, chemicals, psychological or organizational constraints (characterized by hazard codes) to which the patient was exposed during his or her occupational activity (characterized by occupation and activity sector codes). The hazards are coded using a hierarchical code owned by the national social security organism (Caisse Nationale d'Assurance Maladie, CNAM), the patient's professional activity is coded according to the international classification of occupational type (Classification Internationale Type des Professions, CITP-88), the French equivalent of the International Standard Classification of Occupations (ISCO), and the activity sector is coded according to the French occupational classification (Nomenclature des Activités Professionnelles, NAP-03). Two OHPs may differ either by their pathology and/or by at least one component of the exposure, i.e., hazards, occupation and activity sector. Likewise, several patients can present an identical OHP as well. From 2002 to 2007, the RNV3P database was full of 90,335 OHPs of which 75% were reported associated with one hazard, 17% with two hazards, 5% with three hazards, 2% with four hazards and less than 1% with five hazards.

3. Theory

3.1. Construction of the occupational exposome

To study the complexity of the relationships between diseases and composite occupational exposures and to seek similarities in associated exposures that potentially lead to identical diseases, we developed the concept of the occupational exposome. To process the RNV3P data in the form of a network of relationships, each OHP in the RNV3P database is represented by a node (vertex) $v = (p,e)^{T}$, which is a unique combination of a disease (pathology) "p" and its associated three dimensional composite occupational exposure $e = (h, o, s)^{T}$, characterized by a set of hazards "h", an occupation "o" and an activity sector "s" (Fig. 1). By convention, the hazard vector $h = (h_1, h_2, h_3, h_4, h_5)^{T}$ may comprise from 1 up to 5 distinct hazards that were present in the patient's occupational environment as characterized by "o" and "s", and are suspected or confirmed to be related or cause the disease. Each node (i.e., OHP) is weighted by the total number "w" of identical OHP copies in the database (see Supplementary information Fig. S3). In this way, the initial RNV3P database of OHPs is mapped into

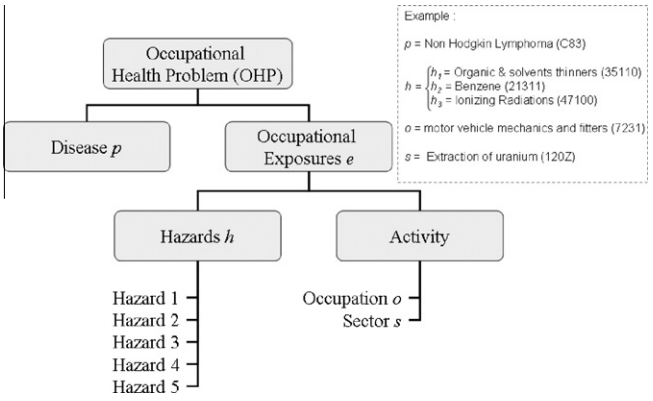

Fig. 1. Structure of an Occupational Health Problem (OHP). Inset: example of an OHP with item codes given between parentheses.

an ensemble of V distinct nodes (each of which being an ensemble of identical OHPs) from which network analyses can now be conducted.

From now, the exposome graph is constructed as follows. Let $C_{ij,\eta}$ (with $0 \leqslant C_{ij,\eta} < 5$) denotes the number of hazards shared by two nodes v_i and v_j ($i, j = 1, 2, \ldots, V$) appearing each at least in η copies in the database. We have, $C_{ij,\eta} = \theta(w_i - \eta) \times \theta(w_j - \eta) \times \sum_{k,l=1}^{5} \delta_{h_{ik},h_{jl}}$, where $\delta_{h_{ik},h_{jl}}$ is the Kronecker symbol, which compares the components h_i and h_j of exposures e_i and e_j, with $\delta_{h_{ik},h_{jl}} = 1$ if $h_{ik} = h_{jl}$, and $\delta_{h_{ik},h_{jl}} = 0$ if $h_{ik} \neq h_{jl}$, and the product $\theta(w_i - \eta) \times \theta(w_j - \eta)$ ensures that the numbers of OHP copies in nodes v_i and v_j are each greater than or equal to a threshold η (with $\eta \geqslant 1$), where $\theta(z) = 1$ if $z \geqslant 0$ and $\theta(z) = 0$ for $z < 0$. Now, we define the connection rule such that two nodes v_i and v_j are connected if their associated $C_{ij,\eta} \geqslant D$, with $1 \leqslant D \leqslant 4$, and not connected otherwise. The D value is comprised between 1 and 4 because any node may be associated with up to five hazards and, therefore, two connected nodes may share D hazards. Thus, for a given D, we define the D_η – occupational exposome as an undirected network characterized by the $V \times V$ adjacency matrix $\mathbf{A} = (A_{ij})$ between nodes v_i and v_j with the elements $A_{ij} = 1$ if $C_{ij,\eta} \geqslant D$ and $A_{ij} = 0$ otherwise. The choice of using the above connection rule for the network construction rather than similarity degree between nodes or other descriptors is dictated by the interest in identifying common or shared elements of the occupational exposure.To summarize, the D_η – occupational exposome is an undirected (weighted) network of OHPs having in common at least D elements of the exposure and where each OHP appears at least η times in the database (see Supplementary information). Such a network is represented by the graph $G = \{W,V,L,D,\eta\}$ made up of a set of W OHPs described by V distinct nodes connected between them by L heterogeneous links and characterized by the adjacency matrix defined by $A_{ij} = 1$ if $C_{ij,\eta} \geqslant D$ and by $A_{ij} = 0$ otherwise, with $i, j \in \{1, \ldots, V\}$. Links in the exposome are heterogeneous because in the same network a node may both share more than one given hazards with a node and share other hazards with another node. The η and D can be regarded respectively as quantitative and qualitative control parameters of the exposome architecture as η controls both the node weight and degree and D, impacting the node degree, allows zooming in or zooming out the network. The probability that two nodes of the network taken at random are connected is given by the density, $d = 2L/V(V-1)$.

3.2. Identification of occupational exposure groups

To structure and organize the occupational exposome in terms of informative motifs or sub-sets carrying information on occupational exposures related to a disease, we introduce the notion of the D_η – occupational exposure group defined as a cluster of nodes (OHPs), in the D_η – occupational exposome, sharing exactly D identical hazards ($1 \leqslant D \leqslant 4$, by definition). Let $\{h_\alpha, \alpha = 1, 2, \ldots, H\}$ be the ensemble or list, of size H, of distinct hazards "h_α" in all nodes of the D_η – occupational exposome, and $\mathbf{B} = (B_{\alpha i})$ the $H \times V$ matrix defined as $B_{\alpha i} = 1$ if $\sum_{k=1}^{5} \delta_{h_{ik},h_\alpha} = 1$, i.e. if one component of the hazard h_i, in the node v_i, is equal to h_α, and $B_{\alpha i} = 0$ if $\sum_{k=1}^{5} \delta_{h_{ik},h_\alpha} = 0$. The 1_η – occupational exposure group (or simple exposure group), coined "h_α" of the name of the hazard under consideration, is given by the ensemble $g_1(h_\alpha)$ of nodes all satisfying $B_{\alpha i} = 1$, i.e. $g_1(h_\alpha) = \{\cup v_i | B_{\alpha i} = 1, i = 1, 2, \ldots, V\}$. The size or number of nodes in $g_1(h_\alpha)$ is, $n_\alpha = \sum_{i=1}^{V} B_{\alpha i}$, with frequency $q_\alpha = \sum_{i=1}^{V} w_i B_{\alpha i} / \sum_{i=1}^{V} w_i$) normalized to one. High order ($D > 1$) occupational exposure groups are constructed from the intersection of D simple exposure groups as, $g_D(h_{\alpha_1}, h_{\alpha_2}, \ldots, h_{\alpha_D}) = g_1(h_{\alpha_1}) \cap g_1(h_{\alpha_2}) \cap \cdots \cap g_1(h_{\alpha_D})$, or equivalently, $g_D(h_{\alpha_1}, h_{\alpha_2}, \ldots, h_{\alpha_D}) = \{\cup v_i | B_{\alpha_1 i} \times B_{\alpha_2 i} \times \cdots \times B_{\alpha_D i} = 1, i = 1, 2, \ldots, V\}$, with normalized frequencies $q_{\alpha_1,\alpha_2,\ldots,\alpha_D} = \sum_{i=1}^{V} w_i [B_{\alpha_1 i} \times B_{\alpha_2 i} \times \cdots \times B_{\alpha_D i}] / \sum_{i=1}^{V} w_i$. Any node with D hazards long is likely to belong to at most equivalently, $2^D - 1 - \delta_{D,5}$, distinct occupation exposure groups. As a consequence, the standard disease–exposure association (or simply, disease–hazard association) is now replaced by the disease–exposure group association where the identified occupational exposure group is considered as the most likely related candidate in the relationships between the disease and a set of exposure situations. Accordingly, the occupational exposome is described as an expansion in terms of occupational exposure groups as D_η — occupational exposome $= \sum_{j=D-1}^{4} m_j g_j(\{h\}) \equiv (m_{D-1}, m_D, \ldots, m_4)^T$, where m_j is the number of j_η – occupational exposure groups. Beyond the nodes, what main matters in exposure groups are hazards which make up situations of occupational exposure. For $j = 0$, m_0 is the number of unconnected nodes and $g_0(\{h\})$represents a single node with hazards not belonging to the list of shared hazards. As both nodes and links never disappear in the growing RNV3P data network, the ensemble of non-shared hazards in the 0_η – occupational exposure groups required specific attention as newly incorporated nodes will be likely to create new connections between any nodes of the network and, therefore, create new exposure groups.

4. Results

To illustrate what an exposome looks like and how this approach can be used to gain information on OHPs, we consider the case of Non-Hodgkin Lymphomas (NHLs), a cancer whose incidence has been increasing since 1970s and whose risk factors are not yet well known [12]. In the present state of knowledge, NHLs are hematological malignancies involving a proliferation of lymphoid cells which tend to infiltrate all organs (gastrointestinal tract, skin, bones, etc.) in an organism. Identified causes of NHLs mainly include immune system abnormalities and factors of infectious origin, although environmental factors, especially occupational exposures, are suspected to increase the risk of NHLs. Indeed, the literature suggests that exposure to substances such as dioxins, pesticides, benzene or organic solvents increase the risk of NHL [13–15]. As the level of proof issue is still very low, this disease is not yet considered in France as a compensated disease.

This knowledge can be extracted from the RNV3P database which, to date, comprises a total of 77 NHL patients or OHPs with a male/female ratio of 4.5 (63/14), mean age of 52.3 years (52 and 53.7 years for male and female, respectively) and median age of 54 years for both genders. This NHL sample involves 72 distinct hazard codes (or 54 aggregated codes), 55 different occupation codes and 28 different codes of activity sectors (Supplementary information Tables S3–S5) thus showing the wide spectrum of occupational exposures that can be associated with NHLs.

4.1. Space of Shared hazards between Non-Hodgkin Lymphomas

Prior to the occupational exposome, Fig. 2 displays as a polar graph the information content of each node in terms of total number of hazards per node and number of hazards shared with other nodes. Each node is represented by a radius of length equal to the number of associated hazards (1–5). The entire surface of the graph (dashed area) represents the available space if all the 51 nodes of OHPs were associated with five hazards each. Our sample of NHLs occupies 60% (dark area) of the available space of which 32% of nodes (16/51) are of length one (associated to one hazard) and the remaining 68% (35/51) of nodes are multi-exposed (length > 1), thus indicating the likelihood that NHLs are associated with more than one hazard. Of the 68% (35/51) multi-exposed nodes, more than 80% (29/35) shares at least one reported hazard with other nodes (totalizing 48% (gray area) of the available space).

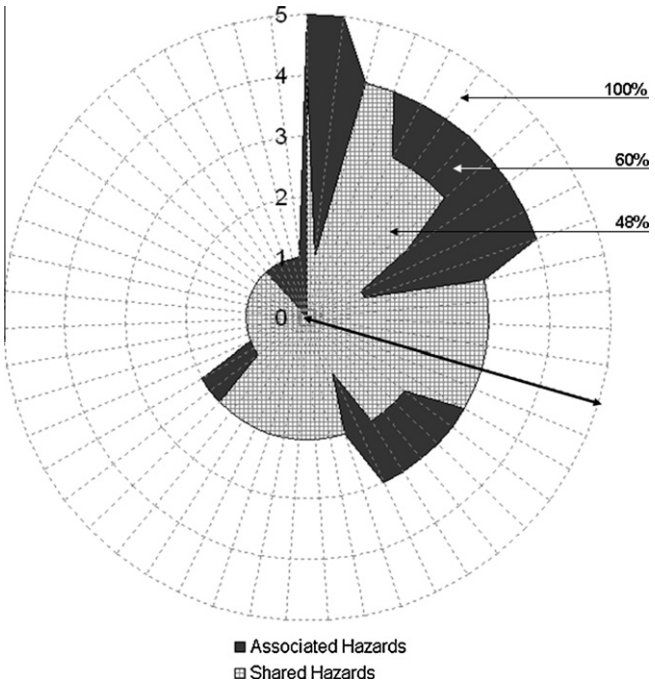

Fig. 2. Distribution of the occupied and shared space among NHLs nodes. The polar graph displays five concentric circles, which represent the five possible associated hazards, with 51 radii each of which corresponding to a node of the NHL sample from the RNV3P database 2002–2007. Shaded gray area represents the space of shared hazards and dark area that of associated hazards. From the 77 NHL OHPs of which nearly half (38/77) are found associated with more than one hazard, 51 distinct nodes are formed; each node representing w OHPs.

4.2. Presentation of Non-Hodgkin Lymphomas exposome

As malignant tumors, NHLs are contained in a node that belongs within the hierarchy of the larger occupational exposome of malignant tumors (Supplementary information Fig. S2 showing the 1_1 and 2_{10} – exposomes). Zooming in that node, Fig. 3 shows the 1_1 – exposome of NHLs generated using NetDraw software [16]. Of the 51 nodes, five are singletons (not connected) as they do not share any hazards with any other nodes, and 46 nodes form a network showing the structure of the spectrum of exposures reported by physicians related to the NHLs. Other underlying dimensions of the exposome structure, the occupation, activity sector and timing of the record entry in the database, are also superimposed for a few nodes in Fig. 3. Using different values of η and D will lead to a different exposome profile as the number of nodes and the exposome density both decrease with η and D (Supplementary information Fig. S2). As an example of variety in situational exposures, the occupation "mechanical engineering technicians" (solid line and open squares in Fig. 3) provided one case every year from 2004 to 2007 from different consultation centers, while the occupation "motor vehicle mechanics and fitters" (solid line and open circles in Fig. 3) reported one case in 2006 and three in 2007 (Supplementary information Table S4) from three different consultation centers. Motor vehicle mechanics belonging to different activity sectors, share similar hazards, in particular benzene as well as other solvents, whereas experts do not report the same hazards for mechanical engineering technicians.

4.3. Topology of the Non-Hodgkin Lymphomas exposome

Two indicators of interest can be used to characterize the NHL – exposome: the degree k_i of a node v_i, defined as the number of nodes in its nearest neighborhood, and the local clustering coefficient c_i for a node v_i, given by the proportion of links between the nodes within its nearest neighborhood divided by the number of links that could possibly exist between them [17,18]. The degree k and clustering coefficient c for each node and their corresponding distributions for the NHL exposome in Fig. 3 are shown in Fig. 4.

The NHL exposome architecture does not seem to fit with any of the well-known networks such as exponential, Erdos-Rényi, Barabasi-Albert and small world networks [19–22]. However, as in many empirically observed networks including the protein networks [23] and some social networks, the NHL exposome exhibits some features of scale-free networks where both the degree and clustering coefficient distributions follow a power law [24].The first feature of scale-free networks is the commonness of nodes with a degree that greatly exceeds the average. Fig. 4 clearly shows more than half of nodes with a degree greater than the mean $\langle k \rangle = 8.24$ (median = 6, standard deviation = 6.98), and a number of nodes with a clustering coefficient greater than the mean $\langle c \rangle = 0.74$ (median = 0.77, standard deviation = 0.26). Note also that the distribution of degree fits with inverse of square root as $N(k) \sim 1/\sqrt{k}$, a decreasing law different from that found in the literature on scale-free networks. However, one has to keep in mind the yet small size of the NHL network.

The second feature in scale-free networks is that the local clustering coefficient decreases as the node degree increases. This means that the low-degree nodes belong to very dense sub-graphs and these sub-graphs are connected to each other through hubs. As illustrated in Fig. 4, nodes with low k and c bridge clusters of nodes like, for instance, the node "NHL × toluene × viruses × aliphatic aldehydes" which belongs to three clusters (dotted gray box in Fig. 4). Such bridging nodes have an important role in the cohesion and robustness of a cluster structured network such as the NHL exposome. Indeed, removing such hub nodes may lead to disintegration of the network [25,26]. The highest-degree nodes, associated with three hazards on average, are found within the largest cluster of nodes where k, ranging from 10 to 23 links (dotted gray box in Supplementary information Fig. S3), increases as the clustering coefficient decreases. Inspection of the top ten highest-degree nodes (Supplementary information Table S2) already suggests a tentative list of most frequently shared hazards, i.e., of leading occupational exposure groups. In addition, as the clustering coefficient of a node quantifies how close its neighbors are to forming a complete graph or a clique, the highest-degree nodes with a moderate clustering coefficient can be regarded as hubs between occupational exposure groups while those with high clustering coefficient belong to a single clique. Interestingly, we found that the NHL – occupational exposome is an assortative mixing network with an assortative index of $r = 0.321$, indicating that well connected nodes tend to connect one another [27]. As connected nodes in the exposome have in common at least an occupational hazard, a positive assortativeness suggests the occurrence of very popular hazards.

4.4. Occupational exposure groups of Non-Hodgkin Lymphomas

The occupational exposome allows structuring expert knowledge in the form of occupational exposure groups. This supervised learning of individual characteristics enables to focus on a collective scale, as reported in Fig. 3. Indeed, we found that the complexity of the NHL exposome is structured as 1_η — NHL exposome $= (5, 24, 48, 11, 3)^T$, i.e., five unconnected nodes, 24 simple exposure groups (compared to the 54 distinct hazard codes), 48, 11 and 3 lists of 2_1, 3_1 and 4_1 – occupational exposure groups, totalizing 86 occupational exposure groups. These exposure groups constitute, within the framework of the RNV3P, the backbone tree spectrum of the most likely occupational exposures that may be related to NHLs.

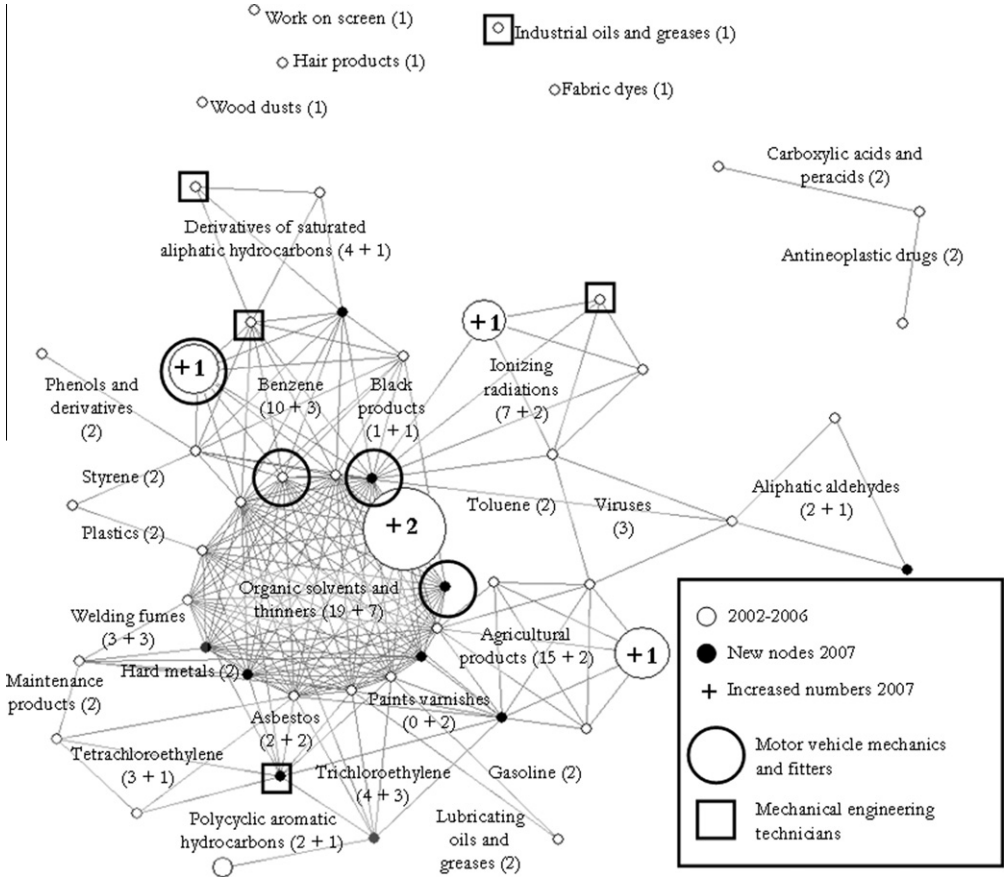

Fig. 3. The 1_1 – exposome of Non-Hodgkin Lymphomas. RNV3P database 2002–2007, with $G_{NHL} = \{W = 77, V = 51, L = 210, D = 1, \eta = 1\}$; W: total number of OHPs, V: number of nodes and L: number of links. A node represents a set of identical NHL OHPs associated with a cortege of 1–5 hazards and contains at least $\eta = 1$ OHP. Nodes are connected if they share at least $D = 1$ hazard. The size of the nodes is proportional to their weight in number of OHPs. Quoted hazard names represent the 1_1 – occupational exposure groups with associated number of OHP content given in parentheses (total number of OHPs in period 2002–2006 + number of OHPs in 2007). There are five unconnected nodes at the top corresponding to the 0_1 – occupational exposure groups.

The spectrum of simple occupational exposure groups (i.e., the most frequently shared hazards) are quoted in Fig. 5 and listed in Fig. 6 with corresponding size and frequency distributions. Both, the size and frequency distributions are similar with the largest (Fig. 5) and most frequent (Fig. 6) group being the "organic solvents and thinners". All top ten highest-degree nodes belong to the most frequent exposure groups although some of their associated hazards are not shared with other nodes (Supplementary information Fig. S1). Fig. 5 also shows the overlaps between exposure groups and that high order exposure groups are often formed from larger and more frequent simple exposure groups, thus implying a high probability of occupational multi-exposure in NHLs. Overlaps between occupational exposure groups indicate that low order exposure groups are embedded in high order ones and/or sharing hazards for exposure groups of the same order >1.

Note that simple exposure groups can be considered as homogeneous cliques since a clique is formally defined as a sub-set of n nodes connected by $n - 1$ links (regardless the nature of links) to other nodes in that clique [28,29]. Thus, because of the possible heterogeneity in the nature of links, the cliques generate "hybrid" exposure groups where links between nodes in the clique may be of different kinds (i.e., the shared hazards are different between nodes in the same clique). As an example of a hybrid exposure group, the "petrol x lubricating oils and greases x organic solvents and thinners" group is made up of three fully connected (through two links) nodes where each of two nodes share a single hazard "petrol", "lubricating oils and greases" and "organic solvents and thinners". Using algorithms to compute a census of all cliques, we found 22 hybrid exposure groups in the 1_1 – NHL occupational exposome.

5. Discussion

We have outlined the theoretical framework of the occupational exposome, defined as a network of OHPs linked by similar occupational exposures, which allows to characterize and to analyze the disease–exposure associations reported in the RNV3P database in the form of a relational network. The main underlying idea of this approach is to investigate characteristics or traits which gather or separate OHPs and therefore occupational exposure. Accordingly, the occupational exposome allows processing the RNV3P data in the form of an OHP network which, in turn, can be studied and characterized as any network and extract interesting and useful properties. In addition, we have structured the occupational exposome in terms of occupational exposure groups which, within the RNV3P framework and in the occupational health context, are informative sub-sets of hazards considered as the backbone tree spectrum of the occupational exposures which are potentially related to the disease under consideration. With this approach OHP data can be analyzed at the scale of a single OHP or at that of an ensemble of OHPs as well.

For the sake of simplicity and of arguments, we have presented implementation of the exposome approach only at the scale of a

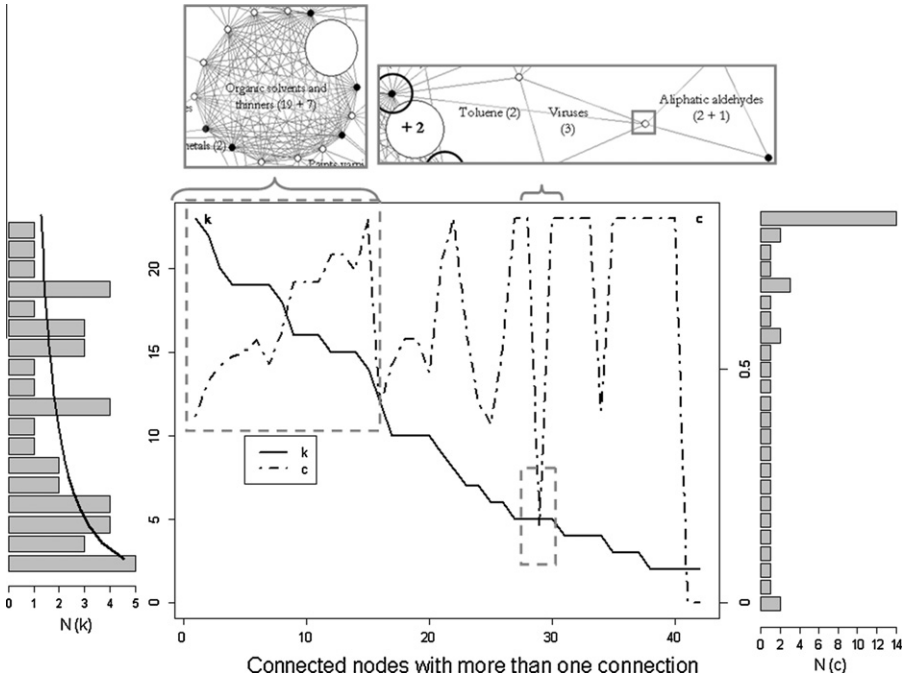

Fig. 4. Distribution of connectivity degree and clustering coefficient. The main frame shows the number of links per node or degree k (left *y*-axis) and the corresponding clustering coefficient *c* (right *y*-axis) versus the nodes with more than one connection in the 1_1 – exposome of NHLs. The insets in the top zoom on the corresponding parts of the NHL exposome in Fig. 3. The histograms on the left and right represent the distributions $N(k)$ of degree k and $N(c)$ of clustering coefficient c, respectively. The solid line through the histogram represents the best fit with the power law, $N(k) \sim k^{-0.52}$.

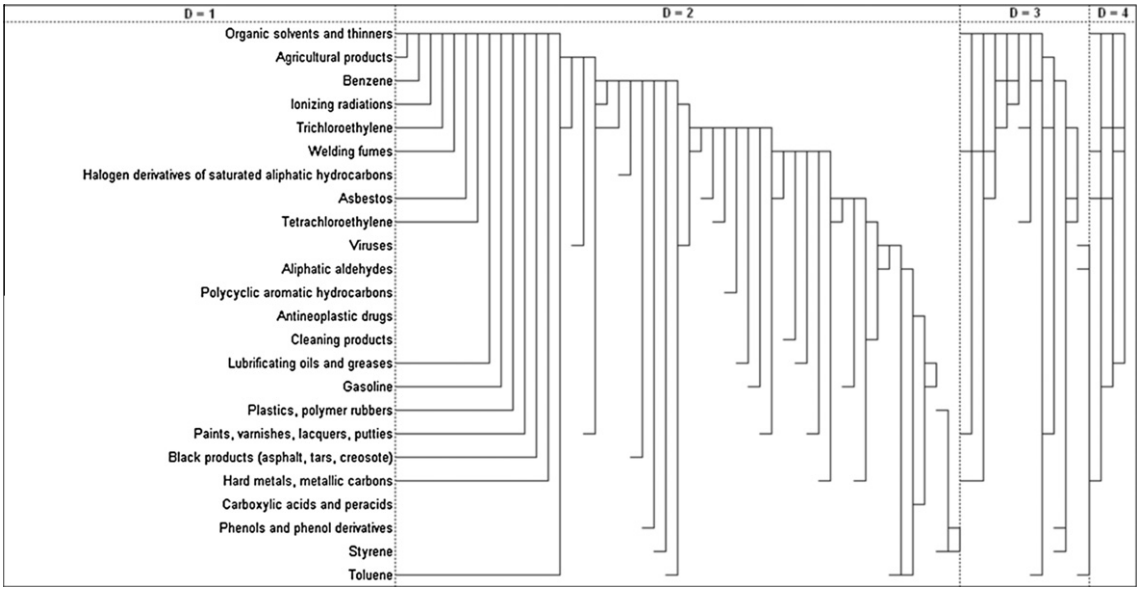

Fig. 5. Distribution of 24 shared hazards according to different multi-exposure levels (i.e. number of shared hazards). List of the 86 exposure groups of NHLs comprising 24 simple exposure groups ($D = 1$), and 48, 11 and 3 high order exposure groups 2_1, 3_1 and 4_1, respectively, RNV3P database 2002–2007. For $D = 1$, the exposure groups are sorted in decreasing order of their frequency (see Fig. 6) from the top to the bottom. For $D > 1$, each exposure group (constructed from D simple exposure groups) is represented by a vertical solid line with D branches each of which pointing on an associated simple exposure group.

single OHP, although the occupational exposome for several OHPs (malignant tumors) is given in Supplementary information (Fig. S2). As illustrated with the NHLs sample (see Figs. 3-6), the 1_1 – occupational exposome of the NHL is a "scale-free" like assortative mixing low dense ($d = 0.165$) network made up of 51 nodes with mean degree and mean clustering coefficient of 8.24 and 0.74, respectively, and structured in 86 occupational exposure groups plus five unconnected nodes corresponding to the 0_1 – occupational exposure groups (see Figs. 5 and 6). These non-shared hazards in the 0_1 – occupational exposure groups required specific attention as they are likely to create new occupational exposure groups when incorporated new nodes during the growing of the network. In addition, beside the small size of the NHL's sample, it would be informative to understand the underlying process leading to the inverse square root power law for the degree distribution in the NHL occupational network.

Although the small size of the NHL sample does not allow reliable fitting, the results found from the NHL occupational exposome

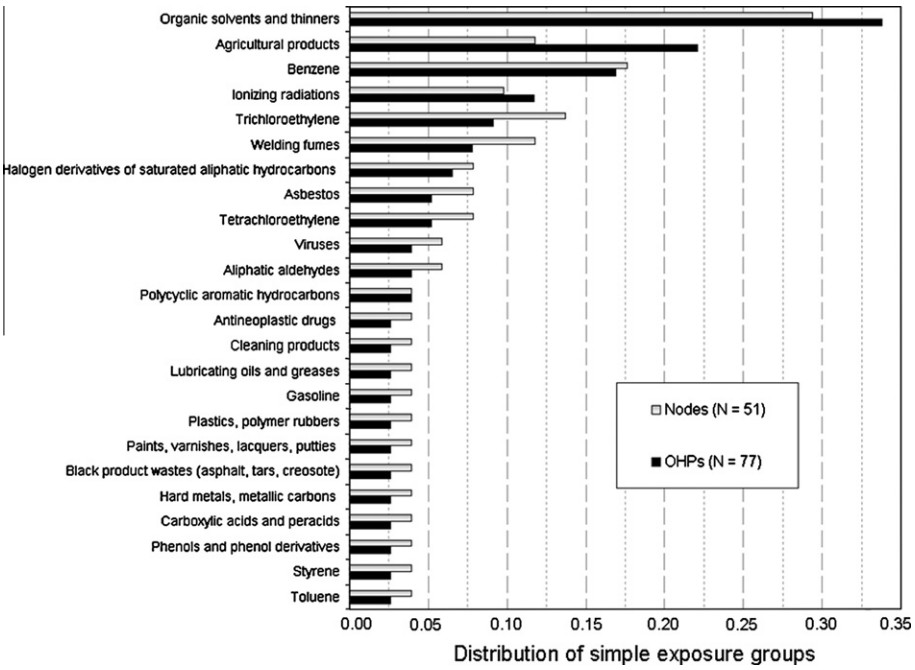

Fig. 6. Frequency of simple occupational exposure groups. Frequency q_α (dark histogram) of the h_α simple exposure groups of the 1_1 – exposome of NHLs, RNV3P database 2002–2007. Light histogram represents proportion of nodes in each exposure group.

can be considered as the state of the knowledge at the date of available data in the RNV3P. These findings turn out to be in coherence with the state of the knowledge on occupational and environmental risk factors of NHL. Indeed, recalling that the RNV3P database is a growing database implies that both nodes and links between nodes never disappear in the occupational exposome. The possible modifications in such a growing network will be to increase the weight of existing nodes, create new connections between already existing nodes, add new nodes and create connections between already present and new nodes. These changes occurring during the growing process of the network may affect some properties (like distributions of degree and clustering) of the occupational exposome but not destroy the structure in occupational exposure groups and rather increase the total number of occupational exposure groups with changes in their relative frequencies.

The approach that we are currently developing finds already immediate applications in addressing issues of the RNV3P. Indeed, organizing the ensemble of OHPs by mean of exposome serves as the starting point for further analyzes of the data. And, in contrast to analytic epidemiology that traditionally aims to test hypotheses advanced beforehand, the occupational exposome approach outlined above is intended to be used for (1) scheduled surveillance of identified and targeted OHP groups (or disease–occupational exposure associations) and for (2) prospective surveillance of the evolving network for detecting emergent events (like growth of already existing and/or appearance of new nodes and/or occupational exposure groups) leading to generation of new hypotheses in relationships between disease and occupational exposure that can be investigated by further epidemiological studies.

Finally, in the new version of the RNV3P information system, the surveillance approach will include three complementary analysis methods allowing generation of pre-alert signals: automatic data-mining procedure based on statistical tests of disproportionate as used in pharmacovigilance which singles out potentially emergent disease–exposure couples, occupational exposome approach applied both on pre-selected single diseases and classes of diseases and, clinical alerts procedure. It is expected that the combined strength of the three approaches to be altogether synergic for surveillance and detection of emerging OHPs: a pre-alert signal generated by one method may be also investigated by the others.

6. Conclusion

The occupational exposome approach opens up new horizons and we have hardly begun to explore the wide possibilities of explanation that it offers. The occupational exposomes could be analyzed at several levels: at the level of a single disease (as illustrated above for the NHLs) or at the level of several diseases taking into account one, two or three dimensions of the occupational exposure (i.e., hazards, occupation and activity). Each scale of analysis has a corresponding occupational exposome, providing knowledge on the multi-exposure potentially related to the disease. Analysis at the level of several different diseases would result in an interlocking, hierarchical structure of exposomes. An additional interesting feature is that as the RNV3P is a growing database the exposome evolves over time as it has been illustrated in the exposome of NHLs without further analysis. In the framework of monitoring and surveillance over time, one may thus study trends in time of the exposomes as the dynamic of evolving and growing networks.

In prospective, for instance, one may make use of the underlying dynamic like in preferential attachment model (in which node links can be weighted depending on the number of nodes connected to it, i.e., the node degree) for generating expected occupational exposomes, and thus expected OHPs. Finally, this approach could be further applied to a global analysis of health problems that would include in addition to the RNV3P database other types of data, variables and descriptors such as, for instance, those used in the diseasome and the Wild like exposome, or approach employed by Patel et al. [30]. Such a step forward greatly exceeds the already complex framework of the RNV3P and necessitates going into and/or combining additional information and other databases.

Acknowledgments

The authors thank the French Agency for Food, Environmental and Occupational Health and Safety (Agence nationale de sécurité sanitaire, ANSES) for supporting this work, the national health insurance organization (Caisse Nationale d'Assurance Maladie, CNAM) for funding of the consultations and participation in funding of the RNV3P, the engineers of the regional branches of the national health insurance organization (Caisses Régionales d'Assurance Maladie, CRAM), Sylvette Liaudy for her help with bibliography, Lynda Larabi for her contribution to the processing and quality control of data, the staff of the occupational health consultation centers who supply the RNV3P with data (C. Doutrellot-Philippon (Amiens), D. Penneau-Fontbonne Y. Roquelaure (Angers), I. Tahon (Besançon), P. Brochard, C. Verdun-Esquer (Bordeaux), J.D. Dewitte (Brest), M. Letourneux (Caen), M.F. Marquignon (Cherbourg), A. Chamoux, L. Fontana (Clermont-Ferrand), J.C. Pairon (Créteil), H.J. Smolik (Dijon), J. Ameille, A. d'Escatha (Garches), A. Maitre, E. Michel (Grenoble), A. Gislard (Le Havre), P. Frimat, C. Nisse (Lille), D. Dumont (Limoges), A. Bergeret, J.C. Normand (Lyon), M.P. Le Hucher-Michel (Marseille), C. Paris (Nancy), D. Dupas, C. Geraut (Nantes), D. Choudat (Paris – Cochin), R. Garnier (Paris – Fernand Widal), D. Leger (Paris – Hotel-Dieu), E. Ben-Brik (Poitiers), F. Deschamps (Reims), A. Caubet, C. Verger (Rennes), J.F. Caillard, J.G. Gehanno (Rouen), D. Faucon (Saint-Etienne), A. Cantineau, (Strasbourg), J.M. Soulat (Toulouse), G. Lasfargues (Tours), the sentinel physicians of the occupational health services for recording and transmitting incident occupational health reports, and Nina Crowte for assistance in the translation from French of the manuscript.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2011.02.010.

References

[1] Chen Y, Turner S, Hussey L, Agius R. A study of work-related musculoskeletal case reports to the Health and Occupation Reporting network (THOR) from 2002 to 2003. Occup Med (Lond) 2005;55:268–74.
[2] Turner S, Lines S, Chen Y, Hussey L, Agius R. Work-related infectious disease to the Occupational Disease Intelligence Network and the Health and Occupation Reporting network in the UK (2000–2003). Occup Med (Lond) 2005;55:275–81.
[3] Walsh L, Turner S, Lines S, Hussey L, Chen Y, Agius R. The incidence of work-related illness in the UK health and social work sector: the Health and Occupation Reporting network 2002–2003. Occup Med (Lond), vol. 55. p. 262–7.
[4] Bonneterre V, Bicout DJ, Larabi L, Bernardet C, Maitre A, Tubert-Bitter P, et al. Detection of emerging diseases in occupational health: usefulness and limitations of the application of pharmaco-vigilance methods to the database of the French National Occupational Dise ase Surveillance and Prevention network (RNV3P). Occup Environ Med 2008;65:32–7.
[5] Bonneterre V, Faisandier L, Bicout D, Bernardet C, Piollat J, Ameille J, et al. For RNV3P. Programmed health surveillance and detection of emerging diseases in occupational health: contribution of the French national occupational disease surveillance and prevention network (RNV3P). Occup Environ Med 2009;67:178–86.
[6] Faisandier L, Bonneterre V, De Gaudemaris R, Bicout DJ. Development of a statistical method to detect emerging events: application to the French national occupational disease surveillance and prevention network (Réseau National de Vigilance et de Prévention des Pathologies Professionnelles, RNV3P) (Translated from French). Epidémiol et Santé Anim 2007;51:111–8.
[7] Faisandier L, De Gaudemaris R, Bicout DJ. Occupational Health Problem Network: The Exposome. <http://arxiv.org/abs/0907.3410>, unpublished results.
[8] Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers Prev 2005;14:1847–50.
[9] Barabási AL. Network medicine – from obesity to the "diseasome". New Engl J Med 2007;357(Suppl 4):404–7.
[10] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. Proc Natl Acad Sci USA 2007;104:8685–90.
[11] Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. New Engl J Med 2007;357:370–9.
[12] Müller AM, Ihorst G, Mertelsmann R, Engelhardt M. Epidemiology of non-Hodgkin's lymphoma (NHL): trends, geographic distribution, and etiology. Ann Hematol 2005;84:1–12.
[13] Alexander DD, Mink PJ, Adami HO, Chang ET, Cole P, Mandel JS, et al. The non-Hodgkin lymphomas: a review of the epidemiologic literature. Int J Cancer 2007;120(Suppl 12):1–39.
[14] Fabro-Peray P, Daures JP, Rossi JF. Environmental risks factors for non-Hodgkin's lymphoma: a population-based case-control study in Languedoc-Roussillon, France. Cancer Causes Control 2001;12(Suppl 3):201–12.
[15] Viel JF, Arveux P, Bavarel J, Cahn JY. Soft-tissue sarcoma and non-Hodgkin's lymphoma clusters around a municipal solid waste incinerator with high dioxin emission levels. Am J Epidemiol 2000;152(Suppl 1):13–9.
[16] Borgatti SP, Everett MG, Freeman LC. Ucinet for windows: software for social network analysis. Harvard (MA): Analytic Technologies; 2002.
[17] Newman ME, Watts DJ, Strogatz SH. Random graph models of social networks. Proc Natl Acad Sci USA 2002;99(Suppl 1):2566–72.
[18] Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature 1998;393(6684):440–2.
[19] Barabási AL, Albert R. Emergence of scaling in random networks. Science 1999;286:509–12.
[20] Barabási AL. The architecture of complexity. IEEE Contr Syst Mag 2000;27(4):33–42.
[21] Palla G, Derenyi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature 2005;435:814–8.
[22] Park J, Barabási AL. Distribution of node characteristics in complex networks. Proc Natl Acad Sci USA 2007;104:17916–20.
[23] Yook SH, Oltvai ZN, Barabasi AL. Functional and topological characterization of protein. Interact networks 2004;4:928–42.
[24] Albert R, Barabási AL. Statistical mechanics of complex networks. Rev Mod Phys 2002;74:47–97.
[25] Vallabhajosyula RR, Chakravarti D, Lutfeali S, Ray A, Raval A. Identifying hubs in protein interaction networks. PLoS One 2009;4(4):e5344.
[26] He X, Zhang J. Why do hubs tend to be essential in protein networks? PLoS Genet 2006;2(6):e88.
[27] Newman MEJ. Assortative mixing in networks. Phys Rev Lett 2002;89:208701.
[28] Adamcsek B, Palla G. CFinder: locating cliques and overlapping in modules in biological networks. Bioinformatics 2006;22(8):1021–3.
[29] Onnela J-P, Saramäki J, Hyvönen J, Szabo G, Lazer K, Kertesz J, et al. Structure and tie strengths in mobile communication networks. Proc Natl Acad Sci USA 2007;104(18):7332–6.
[30] Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus. PLoS One 2010;5(5):e10746.

Glossary

RNV3P: (Réseau National de Vigilance et de Prévention des Pathologies Professionnelles) the National Occupational Disease Surveillance and Prevention Network is a network of physician specialists from 30 occupational pathology centers of university hospitals in metropolitan France. Any patient diagnosed in these centers as having a disease potentially related to occupational exposures is recorded as an Occupational Health Problem (OHP) in the national database

OHP: corresponding to a patient case in the RNV3P database an Occupational Health Problem is an association of a disease (or pathology) and a composite occupational exposure (hazards occupation and/or activity sector) of which one or several hazards are potentially causative

D_η- occupational exposome: undirected network of OHPs having at least D components of the occupational exposure in common and in which each OHP appears at least η times in the database

D_η- occupational exposure group: cluster of OHPs in the $\boldsymbol{D_\eta}$ – occupational exposome sharing exactly $\boldsymbol{D}$ identical hazards of the occupational exposure

NHL- Non-Hodgkin Lymphoma