Detection of emerging diseases in occupational health: usefulness and limitations of the application of pharmacosurveillance methods to the database of the French national occupational disease surveillance and prevention network (RNV3P)

Vincent Bonneterre,¹ ² Dominique J Bicout,¹ Linda Larabi,² Cyril Bernardet,² Anne Maître,¹ Pascale Tubert-Bitter,³ Régis de Gaudemaris¹ ²

ABSTRACT
Objective: To test data mining methods used in pharmacosurveillance in order to identify potential emerging disease–nuisance associations in the national occupational disease surveillance and prevention network (RNV3P) database.

Methods: Proportional reporting ratios (PRR) used in pharmacosurveillance were applied to detect disproportional reporting of disease–nuisance associations which are not compensated by the national social security system.

Results: The 24 785 reports of the RNV3P were grouped into 1344 different disease–nuisance associations reported more than twice, of which 422 did not give entitlement to compensation by the social security system. Among these associations, 162 were potentially emergent and generated a signal, of which eight associations involve cancer.

Conclusion: This work is the first stage of an exploratory investigation submitting the questions raised to experts and involving participants in the network in reflection on the hypotheses generated.

The detection of new work-related diseases is an important public health issue, especially as large numbers of new processes and products are regularly introduced. The Réseau National de Vigilance et Prévention des Pathologies Professionnelles (RNV3P) was created in 2001 in France to monitor occupational and environmental diseases and, in return, to allow participating centres, when confronted with an unusual potentially work-related problem, to search the RNV3P database for similar cases reported elsewhere in France.¹ ³ One of its aims is to develop statistical methods or models to detect potential emerging associations between diseases and nuisances. Since 2001, this network has carried out systematic and standardised recording of all occupational health reports originating from the 29 consultation centres for occupational diseases of university hospitals in metropolitan France, to which patients are referred by occupational physicians, hospital practitioners, or general practitioners for diagnosis of the occupational nature of their disease.

The aim of this work is to test the value and limitations of the application, to the data in the RNV3P database, of safety data mining methods used in pharmacosurveillance,⁴ ⁵ and in particular proportional reporting ratios (PRR). The goal is to detect previously unrecognised disease–nuisance associations (called “associations”) which do not yet give entitlement to compensation and which generate a signal with these methods. These associations are likely to reveal a new risk for workers’ health, so in this paper they will be termed “potentially emerging associations”. While this approach cannot affirm a causal relationship between the nuisance and the disease observed, it may generate new hypotheses which can be submitted to experts for evaluation.

METHODS
Data analysed concern the occupational health reports collected by the RNV3P database covering the years 2001–2005. The quality of the data is optimised by use of a computerised tool for entering, coding and monitoring the coherence of the information, by annual training in quality assurance for users, by the creation of a national coding group as well as by centralised quality control. Each occupational health report is a structured expert clinical report whose principal coded items are: principal disease and co-morbidities (ICD-10), principal nuisance and four other possible nuisances (INRS-CNAM), professional position (ISCO-88, edited by ILO) and sector of professional activity (NAF, edited by INSEE). These national or international codes are graded and can be analysed at different degrees of precision. An association is considered as recognised, and cannot therefore be emergent, if it potentially gives entitlement to compensation (“compensated association”).

The data mining methods used in pharmacosurveillance are independent of data external to the base. They analyse drug–event pairs which are the equivalent of the disease–occupational nuisance pairs included in the RNV3P. The structure of the disease-occupational nuisances matrix of the RNV3P is similar to that of the adverse drug event matrices used by the principal pharmacosurveillance databases (MEDWATCH in the USA, LAREB in the Netherlands, VIGIBASE for the WHO): they are “empty” matrices, consisting mainly of cells whose content is nil, with a possible reported percentage of pairs of about 1%. For these
reasons, as a preliminary exploratory approach we have chosen to use pharmacosurveillance methods to analyse the RNV3P data.

These methods generate a statistical signal when there is a discrepancy between the true number of a disease–nuisance pair reported in the RNV3P database and the expected number for this same pair in the database. For each of the pairs, these methods generally use a 2×2 contingency table in which the number of reports is entered (table 1).

Each measurement of associations uses a different probability model to describe the distribution of the number of reports. There are few published, large-scale, systematic comparisons of the data-mining methods currently used for pharmacosurveillance. However, the results of the different methods tend to converge when the number of reports increases for a given pair. The method tested here is “proportional reporting ratios” (PRR), used by the British Medicine Control Agency (MCA) to analyse the data of the ADROIT (Adverse Drug Reactions On-line Information Tracking) pharmacosurveillance database. The equation which gives the value of PRR for a given pair is

\[
PRR = \frac{a/(a+b)}{c/(c+d)}.
\]

This is similar to the relative risk used in epidemiology. The PRR is equal to the ratio of the probability \(a/(a+b)\) of presenting, within the database, the target disease if exposed to the target nuisance, over the probability \(c/(c+d)\) of presenting this disease if exposed to any nuisance, with the exception of the target nuisance. Two signal generation criteria have been proposed with PRR: the first (PRR1) uses three conditions: (number of cases observed : a to any nuisance, with the exception of the target nuisance. Two-by-two contingency table

\[
\begin{align*}
\text{Reports with the target disease} & \quad \text{Reports with other diseases} & \quad \text{All diseases} \\
\text{Reports with the target disease} & \quad a & \quad b & \quad a+b \\
\text{Reports with other nuisances} & \quad c & \quad d & \quad c+d \\
\text{All nuisances} & \quad a+c & \quad b+d & \quad N
\end{align*}
\]

Table 1  Two-by-two contingency table

Figure 1  Distribution of the number of compensated and non-compensated disease–nuisance associations reported from 1 to more than 50 times.

\[
CI = \exp(ln(PRR) ± 1.96SE(lnPRR)), \quad \text{with } SE(lnPRR) = \frac{1}{a+b+c+d}/\frac{1}{a}\frac{1}{b}\frac{1}{c}\frac{1}{d}^{\frac{1}{2}}.
\]

A signal is generated if the lower limit of the 95% CI is greater than unity (LI 95(PRR) > 1). According to PRR1 and PRR2 definitions, PRR1 cases are expected to be compensated or non-compensated disease–nuisance associations reported from 1 to more than 50 times.

\[
\chi^2 \text{ Yates 1 dof } \chi^2 \text{ Yates 2 with Yates correction. These two signal}
\]

The 422 non-compensated associations reported more than twice accounted for 5103 reports. Some are well known to clinicians and are very frequently observed, such as psychiatric disorders related to psychosocial and organisational stress (3770 reports).

\[
\chi^2 \text{ Yates 1 dof } \chi^2 \text{ Yates 2 with Yates correction. These two signal}
\]

The 422 non-compensated associations reported more than twice accounted for 5103 reports. Some are well known to clinicians and are very frequently observed, such as psychiatric disorders related to psychosocial and organisational stress (3770 reports).

\[
\chi^2 \text{ Yates 1 dof } \chi^2 \text{ Yates 2 with Yates correction. These two signal}
\]

The 24 785 reports in the RNV3P database gave 3830 different disease–nuisance associations, of which only 47% corresponded to associations giving entitlement to compensation. Figure 1 presents the distribution of the number of compensated and non-compensated associations reported from 1 to more than 50 times. Only one-third of associations were reported more than once (1344 associations); however, these accounted for 59% of all 22 299 reports in the database. Two-thirds of the associations reported at least twice (n = 922) gave entitlement to compensation (table 2), a total of 17 196 reports. The compensated diseases most frequently encountered were related to asbestos (20 different associations, 5485 reports), dermatitis (149 associations, 2485 reports), bone and muscle disorders related to repetitive trauma and load carrying (70 associations, 2284 reports), asthma (132 associations, 1774 reports), occupational deafness (8 associations, 874 reports) and rhinitis (90 associations, 744 reports).

The 422 non-compensated associations reported more than twice accounted for 5103 reports. Some are well known to clinicians and are very frequently observed, such as psychiatric disorders related to psychosocial and organisational stress (3770 reports).

\[
\chi^2 \text{ Yates 1 dof } \chi^2 \text{ Yates 2 with Yates correction. These two signal}
\]

The 24 785 reports in the RNV3P database gave 3830 different disease–nuisance associations, of which only 47% corresponded to associations giving entitlement to compensation. Figure 1 presents the distribution of the number of compensated and non-compensated associations reported from 1 to more than 50 times. Only one-third of associations were reported more than once (1344 associations); however, these accounted for 59% of all 22 299 reports in the database. Two-thirds of the associations reported at least twice (n = 922) gave entitlement to compensation (table 2), a total of 17 196 reports. The compensated diseases most frequently encountered were related to asbestos (20 different associations, 5485 reports), dermatitis (149 associations, 2485 reports), bone and muscle disorders related to repetitive trauma and load carrying (70 associations, 2284 reports), asthma (132 associations, 1774 reports), occupational deafness (8 associations, 874 reports) and rhinitis (90 associations, 744 reports).

The 422 non-compensated associations reported more than twice accounted for 5103 reports. Some are well known to clinicians and are very frequently observed, such as psychiatric disorders related to psychosocial and organisational stress (3770 reports).

\[
\chi^2 \text{ Yates 1 dof } \chi^2 \text{ Yates 2 with Yates correction. These two signal}
\]

The 24 785 reports in the RNV3P database gave 3830 different disease–nuisance associations, of which only 47% corresponded to associations giving entitlement to compensation. Figure 1 presents the distribution of the number of compensated and non-compensated associations reported from 1 to more than 50 times. Only one-third of associations were reported more than once (1344 associations); however, these accounted for 59% of all 22 299 reports in the database. Two-thirds of the associations reported at least twice (n = 922) gave entitlement to compensation (table 2), a total of 17 196 reports. The compensated diseases most frequently encountered were related to asbestos (20 different associations, 5485 reports), dermatitis (149 associations, 2485 reports), bone and muscle disorders related to repetitive trauma and load carrying (70 associations, 2284 reports), asthma (132 associations, 1774 reports), occupational deafness (8 associations, 874 reports) and rhinitis (90 associations, 744 reports).

The 422 non-compensated associations reported more than twice accounted for 5103 reports. Some are well known to clinicians and are very frequently observed, such as psychiatric disorders related to psychosocial and organisational stress (3770 reports).
difficulties for respiratory disorder. Several associations found as potentially emergent in the network are, at this time, being investigated in the scientific community: trichloroethylene and kidney tumour,91 laryngeal tumour and asbestos,11,12 malignant connective tissue tumour and pesticides,13 lung cancer and paints, organic solvents and thinners,14,15 systemic sclerosis and solvents,16,17 sleep disorders and solvents,18,19 multiple chemical sensitivity and formaldehyde,20 sterility and glycol ether,21 and wood or vegetable dust and sarcoidosis.22 The PRR2 criterion identified all the associations revealed by PRR1, as well as further associations of which 92% were only reported twice in the database. The subgroup of cancerous diseases (table 3) included 72 associations (1613 reports) reported at least twice, comprising 46 compensated associations (1517 reports) and 26 non-compensated associations (96 reports).

The non-compensated associations which generated a signal with the PRR1 criterion are shown in table 4, with the exception of one poorly-coded association which underwent further quality control. The seven complementary associations which generated a signal only with PRR2 and \( \chi^2 \) were all reported twice, except for the association kidney cancer and solvents (four reports).

**DISCUSSION**

Value of the application of pharmacosurveillance methods to the RNV3P

This work is a preliminary step towards automated systematic analysis of RNV3P data in order to detect potentially emerging disease–nuisance associations. The aim is twofold: to reveal, as early as possible, the new occupational risks which can be detected from the network database (hypothesis-generating function) and to emphasise the role of the network in forming expert opinion, as the hypotheses generated are transmitted to experts in the occupational disease centres who assess their pertinence.

The results presented in this paper demonstrate that proportional reporting ratios, as well as identifying compensated associations, have detected some formerly "unrecognised" associations, reported several times by the network, and which originate either from potential toxicological problems (associations where the relationship of cause to effect is not yet established) or from problems of social recognition: work-related health disorders which do not yet give entitlement to compensation. This approach by means of associations has also revealed wide discrepancies in the frequency with which disease–nuisances pairs are reported in the database.

As an illustration, pharmacosurveillance methods could also be used for paring other variables such as disease and activity sector dimension or disease and profession dimension. These new matrices yielded further information. For example, the association of chronic kidney failure (N18) with vehicle maintenance and repair (NAP 502Z) generated a signal with PRR. The hypothesis generated by this signal led to further analysis of these cases. The majority followed a nephritic
syndrome (NOS) cited as a co-morbid disease. They had little chance of attracting the clinicians’ attention as they came from three different consultation centres. Lastly, these cases did not generate signals during analysis of the disease x nuisance or disease x occupation matrices because they had different principal nuisance or occupation codes. It is now the task of the experts to give their opinion on the relevance of this association.

### Inherent limitations of the data

RNV3P suffers from the same limitations as the pharmacosurveillance networks: there is a lack of precise information on the population from which the patients come; a potential link between the patient’s disease and occupational nuisances has to be suggested by a physician, who refers the patient to an occupational health consultation centre in a university hospital; and, lastly, the range of codes used for diseases or nuisances is very broad (synonyms or very closely related concepts) which poses a problem for analysis, as it does in pharmacosurveillance.

With regard to the first point, the aim is not however to calculate the incidence of associations but to look for potentially compensating diseases generating a signal (55% of the associations) which poses a problem for analysis, as it does in pharmacosurveillance.

Lastly, the problem of the broad range of codes interferes with the results. For example, the relatively low proportion of compensated diseases generating a signal (55% of the associations reported at least twice) is due to situations where the observed number of reports is low while the pathology and/or the nuisance are reported a large number of times. This should indicate that the association has become infrequent, but in practice this may also happen because the association is described with different synonymous disease codes or nuisances codes, hence leading the same entity to be greatly increased by several associations. Nevertheless, the compensated diseases generating a signal represent 91% of the reports corresponding to compensated associations reported at least twice (93% for PRR2). This problem of a broad range of codes can be solved by paying attention to coding (standardisation of practices by regular training courses) and also to analysis (grouping of codes). At present, two-thirds of the associations generated in the RNV3P are only reported once, even though the ICD disease code is limited to three digits in order to reach an acceptable compromise between obtaining accurate information and eliminating multiple classification bias related to the large number of disease codes available. An individual approach to the nuisance codes is necessary to assess how they can be grouped without running the risk of losing valuable information in terms of detection. Code groupings will have to be validated by all the experts: a balance must be struck between the need to preserve useful distinctions and the risk of not detecting a signal sufficiently early. However, there is a factor which helps to decrease over time the number of associations reported only once: the network is fed by 5000 new cases every year.

### Inherent limitations of the methods

These statistical methods generate hypotheses which have clinical significance only after they have been evaluated by the

| Table 3 | Distribution of the 72 cancer–nuisance associations reported at least twice between those which do or do not generate a signal with the PRR1 and PRR2 criteria |
|-----------------------------------------------|-------------------------------------------------|-------------------------------------------------|
| **Not compensated** | **Compensated** | **Total** |
| **Association** | **Reports** | **Association** | **Reports** | **Association** | **Reports** |
| **Criterion PRR1** | | | | | |
| Signal | 12% | 3.5% | 36% | 83% | 48% | 86.5% |
| No signal | 9 | 62 | 26 | 1463 | 35 | 1525 |
| Total | 26 | 96 | 46 | 1517 | 72 | 1613 |
| **Criterion PRR2** | | | | | |
| Signal | 22% | 3% | 46% | 82% | 68% | 85% |
| No signal | 16 | 48 | 33 | 1449 | 49 | 1497 |
| Total | 26 | 96 | 46 | 1517 | 72 | 1613 |

Values are ratios in % and number of reports.

### Table 4 | Non-compensated cancer–nuisance associations reported more than twice, and which generate a signal with the PRR1 criterion |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cancer</strong></td>
<td><strong>Nuisance</strong></td>
</tr>
<tr>
<td><strong>Short name</strong></td>
<td><strong>Code</strong></td>
</tr>
<tr>
<td>Larynx</td>
<td>C32</td>
</tr>
<tr>
<td>Lung</td>
<td>C34</td>
</tr>
<tr>
<td>Skin</td>
<td>C44</td>
</tr>
<tr>
<td>Connective tissue</td>
<td>C49</td>
</tr>
<tr>
<td>Prostate</td>
<td>C61</td>
</tr>
<tr>
<td>Kidney</td>
<td>C64</td>
</tr>
<tr>
<td>Kidney</td>
<td>C64</td>
</tr>
<tr>
<td>Non Hodgkin Lymphoma</td>
<td>C85</td>
</tr>
</tbody>
</table>

Legend (same symbols as in the 2 x 2 contingency table represented in table 1): a, reports with target disease and target nuisance; a+c, reports with the target disease; a+b, reports with the target nuisance; LI95%PRR, lower bound of 95% confidence interval of PRR.
experts in relation to other parameters (mechanistic and pathophysiological factors, degree of exposure, duration of exposure, extra-professional risk factors, etc). Several authors warn against “over-confidence in the results of data mining” or “enthusiasm generated by promise of the data mining tools” and remind us that while data mining methods are useful for systematic screening of large databases “to discover hidden patterns of associations or unexpected occurrences”, they must only be used with full awareness of the limitations of these methods and of the databases to which they are applied.

The choice of method and of the criterion of signal generation must be determined so that they do not identify too many associations which would be considered by the experts as false positives. In particular, these methods are highly sensitive to the number of cases reported and to a low number count of the disease and the target nuisance. Therefore, when the number of reports is small, the signal may be very strong even if only two cases are observed. Inversely, certain known and compensated associations do not generate a signal if they are reported relatively few times compared with the number of reports citing either the target disease or the target nuisance. Lastly, a very strong signal generated by a nuisance is likely to attenuate the other associations with the same nuisance, and in the same way a very strong signal with a given disease may attenuate the other associations with the same disease. Other data mining methods used in pharmacosurveillance have been tested on the RNV3P database: other “classical or frequentist forms of disproportionality analysis” (reporting odds ratios or ROR and its derivative Yule’s Q), certain methods using Poisson distribution (Poisson, sequential probability ratio test or SPRT2 methods) and a Bayesian method (Bayesian confidence propagation neural network or BCPNN, also called the information component (IC), WHO or Bate’s method). The sensitivity of these methods certainly differs depending on the number of times the association is reported. However, the above-mentioned tests, applied with the usual thresholds, produced overall about the same proportion of positive signals (with the exception of SPRT2 which is much less sensitive for pairs which are reported only a few times). No extensive study has demonstrated the superiority of any one of these methods compared with another. However, a recent study comparing PRR and the multi-item gamma Poisson shrinker (MGPS, also called empirical Bayes screening or EBS), emphasised the greater stability of the results of the MGPS in situations with small numbers, and so a smaller proportion of false positives. To make data mining tools more effective and to reduce the number of false positives, other factors have been taken into account in analysis, in particular with PRR: “strength of the signal”, its unexpectedness (“whether it is really new”), its “clinical importance (severity and seriousness)”, and its preventable nature (“the potential for preventive measures”). In other words, those given the highest priority are Strong, New, Important and potentially Preventable (SNIP). This list has been completed by the WHO Programme for International Drug Monitoring which proposes supplementary criteria to analyse the results of the BCPNN method, such as a two-fold rise in signal generation over a 3-month period, a positive rechallenge test, etc. This type of filter could well be used in the RNV3P. Another interesting filter could reveal cases from different occupational disease consultation centres (as they may be liable to be missed by experts in each individual centre). Lastly, only the principal nuisance is taken into account in this work, whereas other so-called secondary nuisances may be coded by the experts. Methods taking all nuisances into account are currently being developed.

CONCLUSION
Early detection of new disease–occupational nuisance associations is an important health issue. It cannot be done on the basis on the declarations of compensated occupational diseases, which refer to diseases–nuisance associations already accepted by the experts. The application to the RNV3P data of data mining methods used in pharmacosurveillance has led in particular to the detection of associations which are currently being investigated.

This exploratory investigation has raised clinical queries for submission to the experts and is involving those participating in the network in reflection on the hypotheses generated. It now appears necessary to complete these methods by a modelling approach taking into account all nuisances reported by the expert for each disease and their synergic effects. Other elements which could be taken into account include the degree of imputability attributed by the expert to each nuisance, established from data concerning exposure (duration, dose, chronological relation of symptoms to exposure) and pathophysiological plausibility. When the number of reports allows, further investigation will study the extension of diseases–occupational nuisance associations over time as well as within occupations and sectors of activity.

Acknowledgements: The authors wish to thank all physicians from the occupational health consultation centres of the university hospitals who supply their reports to the RNV3P, as well as the partners and organisers of the network: the Caisse Nationale d’Assurance Maladie (CNAM), the Agence Franc¸aise de Sécurité Sanitaire de l’Environnement et du Travail (A\SSE\T) and the Sociét\'e Franc\'aise de M\'edicine du Travail (SFMT).

Competing interests: None declared

REFERENCES


---

**Date for your diary**

**International Forum on Quality & Safety in Health Care**

22–25 April 2008

Le Palais de Congrès de Paris

Paris, France

http://www.internationalforum.bmj.com

**Why attend?**

- Join over 1000 healthcare professionals from over 40 countries worldwide
- Learn from experienced leaders and improvement experts
- Find out what is current in quality and safety
- Gain new skills and tools for creating change in your organisation
- Take home practical solutions for improvement methods
- Network with like-minded colleagues

This is a premier event for those dedicated to improving quality and safety in healthcare. Our focus is on practical, take-home ideas that will leave you inspired to implement valuable, lasting change on your return to the workplace.

**Register online from January 2008 onwards**

For more information on the forum visit - http://www.internationalforum.bmj.com
Detection of emerging diseases in occupational health: usefulness and limitations of the application of pharmacosurveillance methods to the database of the French national occupational disease surveillance and prevention network (RNV3P)

Vincent Bonneterre, Dominique J Bicout, Linda Larabi, Cyril Bernardet, Anne Maitre, Pascale Tubert-Bitter and Régis de Gaudemaris

*Occup Environ Med* 2008 65: 32-37 originally published online August 29, 2007
doi: 10.1136/oem.2007.033183

Updated information and services can be found at: http://oem.bmj.com/content/65/1/32

These include:

**References**
This article cites 24 articles, 0 of which you can access for free at: http://oem.bmj.com/content/65/1/32#BIBL

**Email alerting service**
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

**Notes**

To request permissions go to: http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to: http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to: http://group.bmj.com/subscribe/