ORIGINAL RESEARCH ARTICLE

Drug Saf 2012; 35 (6): 495-506
0114-5916/12/0006-0495/$49.95/0

Early Detection of Pharmacovigilance Signals with Automated Methods Based on False Discovery Rates

A Comparative Study

Ismaïl Ahmed,[1,2,3] *Frantz Thiessard*,[4,5] *Ghada Miremont-Salamé*,[6,7] *Françoise Haramburu*,[6,7] *Carmen Kreft-Jais*,[8] *Bernard Bégaud*[7,9] and *Pascale Tubert-Bitter*[1,2]

1 INSERM, CESP Centre for Research in Epidemiology and Population Health, U1018, Biostatistics, Villejuif, France
2 Université Paris-Sud, UMRS 1018, Villejuif, France
3 Department of Epidemiology and Biostatistics, Imperial College, London, UK
4 Laboratoire d'Epidémiologie, Statistique et Informatique Médicales (LESIM), Institut de Santé Publique, d'Épidémiologie et de Développement (ISPED), Université Victor Segalen Bordeaux 2, Bordeaux, France
5 Service d'Information Médicale, Centre Hospitalier Universitaire de Bordeaux, Bordeaux, France
6 Centre Régional de Pharmacovigilance, Centre Hospitalier Universitaire de Bordeaux, Bordeaux, France
7 INSERM, U657, Bordeaux, France
8 Département de Pharmacovigilance, AFSSAPS, Saint Denis, France
9 Université de Bordeaux, U657, Bordeaux, France

Abstract

Background: Improving the detection of drug safety signals has led several pharmacovigilance regulatory agencies to incorporate automated quantitative methods into their spontaneous reporting management systems. The three largest worldwide pharmacovigilance databases are routinely screened by the lower bound of the 95% confidence interval of proportional reporting ratio ($PRR_{02.5}$), the 2.5% quantile of the Information Component ($IC_{02.5}$) or the 5% quantile of the Gamma Poisson Shrinker (GPS_{05}). More recently, Bayesian and non-Bayesian False Discovery Rate (FDR)-based methods were proposed that address the arbitrariness of thresholds and allow for a built-in estimate of the FDR. These methods were also shown through simulation studies to be interesting alternatives to the currently used methods.
Objective: The objective of this work was twofold. Based on an extensive retrospective study, we compared $PRR_{02.5}$, GPS_{05} and $IC_{02.5}$ with two FDR-based methods derived from the Fisher's exact test and the GPS model (GPS_{pH0} [posterior probability of the null hypothesis H_0 calculated from the Gamma Poisson Shrinker model]). Secondly, restricting the analysis to GPS_{pH0}, we aimed to evaluate the added value of using automated signal detection tools compared with 'traditional' methods, i.e. non-automated surveillance operated by pharmacovigilance experts.

Methods: The analysis was performed sequentially, i.e. every month, and retrospectively on the whole French pharmacovigilance database over the period 1 January 1996–1 July 2002. Evaluation was based on a list of 243 reference signals (RSs) corresponding to investigations launched by the French Pharmacovigilance Technical Committee (PhVTC) during the same period. The comparison of detection methods was made on the basis of the number of RSs detected as well as the time to detection.

Results: Results comparing the five automated quantitative methods were in favour of GPS_{pH0} in terms of both number of detections of true signals and time to detection. Additionally, based on an FDR threshold of 5%, GPS_{pH0} detected 87% of the RSs associated with more than three reports, anticipating the date of investigation by the PhVTC by 15.8 months on average.

Conclusions: Our results show that as soon as there is reasonable support for the data, automated signal detection tools are powerful tools to explore large spontaneous reporting system databases and detect relevant signals quickly compared with traditional pharmacovigilance methods.

Background

Improving the detection of new drug safety signals has led several national and multinational regulatory agencies to develop and routinely use automated quantitative methods for spontaneous reporting management systems. These recent additions are based on data mining screenings within the spontaneous reporting databases and differ one from another by the chosen disproportionality measures and statistical models. The screenings are processed regularly with every present drug-event pair being assessed. Their common goal is to assist with early identification of potential emerging safety signals. They are intended to provide, prospectively, data lists of statistically detected signals that have to be further investigated before being possibly considered as valid. The first screening methods to be widely used were the Reporting Odds Ratio implemented by the Dutch pharmacovigilance agency (Lareb[1]), the Proportional Reporting Ratio (PRR) used by the Medicines and Healthcare products Regulatory Agency in the UK and at the European Medicines Agency (EMA)[2] and two Bayesian methods, the Information Component (IC)[3,4] for the WHO pharmacovigilance, and the Gamma Poisson Shrinker (GPS) for the US FDA.[5,6] For the GPS, an extension makes it possible to study multi-item associations (multi-item GPS[7]). Currently, none of them can be considered as an undisputed reference. More recently, Bayesian and non-Bayesian False Discovery Rate (FDR)-based methods were proposed that address the arbitrariness of thresholds and allow for a built-in estimate of the FDR,[8] a multiple comparison error criterion defined as the expected proportion of false positives among the list of generated signals.[9,10] These methods were demonstrated in a large simulation study to efficiently minimize the burden of false positive signals compared with the previous methods.[11]

Assessment of pharmacovigilance data mining methods is certainly a challenging issue. If simulation studies are crucial for estimating and comparing statistical properties of methods including classical specificity and sensitivity criteria in an explicitly defined common framework, they need to be complemented by studies within actual reports databases.[11-14] Specificity assessment cannot be realistically addressed by studies based upon actual and not simulated data, as it would require ascertaining true negatives. In an attempt

to tackle the evaluation of sensitivity, retrospective real data application has been performed on some specific examples, in a variety of settings, with a minority of studies involving comparison of the different methods.[5,15-18] Such comparative assessments aim to measure performances of methods with regard to their ability to identify known previously validated signals. However, undisputable selection of a set of reference signals (RSs), i.e. signals considered to be true, is not straightforward. Another difficulty is the large disparity between the numbers of signals generally generated by each method, when the usual detection thresholds are applied. It is sensible to consider that a significant part of these 'signals' could be false positives. However, comparing the various methods proposed with regard to this criterion is unrealistic since an undisputable identification of false positives is not feasible in real-world studies.

To complete their evaluation, the methods should also be compared with traditional, i.e. non-automated, pharmacovigilance detection. Ideally, such evaluation should entail two components, i.e. (i) their ability to detect adverse effects that would not have been pinpointed by experts, and (ii) the ability of the methods to highlight some known drug safety issues in a timely manner compared with traditional pharmacovigilance. While assessing to what extent the automated signal detection methods can improve on the first feature is an unrealistic goal, the present study focuses on the second component. A few retrospective studies have compared the time to detection between automatic signal detection and traditional pharmacovigilance methods.[5,15,19,20] This framework does not rule out the issue of how to define the set of RSs. For all the above-mentioned studies, it relies on labelled adverse effects. It also raises the problem of the choice of a reference index date. As pointed out by Hochberg and Hauben,[19] the latter has no single rigorous definition. It can refer, for instance, as in their study, to the use of simple detection thresholds for report counts or percentages. Another example is given by Alvarez et al.,[20] where the index date for traditional surveillance is defined as the earliest date at which the EMA became aware that a safety signal requiring investigation had been identified.

The present study differs from the above-mentioned studies by defining the set of RSs. It is based on RSs corresponding to investigations launched by the French Pharmacovigilance Technical Committee (PhVTC, 'Comité Technique de Pharmacovigilance') of the AFSSAPS (the French Health Products Safety Agency, 'Agence Française de Sécurité Sanitaire des Produits de Santé') which is in charge of, among other things, coordinating regulatory investigations related to adverse drug reactions. Such an investigation does not necessarily mean that an adverse effect is identified. In fact, only a small proportion of these investigations will eventually lead to an update of the profile of the product's adverse reactions. Nevertheless, these investigations correspond to the earliest signs of potential safety issue awareness of established pharmacovigilance and constitute our RS set. For the latter, we have the date of the launch of an investigation, which is then chosen as the reference index date. It is noteworthy that the set of RSs was not made of safety issues *a posteriori* validated. Another feature is that this set was exhaustive since it was systematically extracted from the 380 alerts launched by the PhVTC during the period 1 January 1996–1 July 2002.[21]

The study is in two parts. First, five automatic signal detection methods were compared as regard to the number of RSs detected as well as their time to detection. The comparison included (i) the three methods used in the three largest worldwide pharmacovigilance databases, i.e. the lower bound of the 95% confidence interval (CI) of PRR ($PRR_{02.5}$), the 2.5% quantile of the information component ($IC_{02.5}$) and the 5% quantile of the GPS (GPS_{05}); (ii) two recent FDR-based methods: midRFET (mid p-values calculated from the Fisher's Exact Test)[9] and GPS_{pH0} (posterior probability of the null hypothesis H_0 calculated from the Gamma Poisson Shrinker model).[10] The second part of the analysis is then restricted to the FDR-based GPS_{pH0} and aimed to evaluate its ability to detect RSs in a timely fashion compared with the traditional detection.

Materials and Methods

French Spontaneous Reporting Database

The data used for this study consisted of all spontaneous reports of adverse drug reactions recorded in the French pharmacovigilance database during the period 1 January 1995–1 July 2002.[22] In order to be analysed with signal detection methods, the whole spontaneous report database was formatted into large contingency tables crossing all the adverse events and drugs reported at least once during a given period of time. The data were coded according to the Anatomical Therapeutic Chemical (ATC) hierarchy,[23] 5th level, chemical subgroup, for the drugs and the Preferred Term granularity of the Medical Dictionary for Regulatory Activities (MedDRA®)[24] for adverse events. The reports are recorded with both the date of occurrence of the event and the date of recording of the case. Only the latter was considered here.

Reference Signals (RSs)

The French pharmacovigilance system is based on 31 regional pharmacovigilance centres, with one of their missions being to monitor and evaluate spontaneous reports of adverse drug reactions made by health professionals of the corresponding area. Every month, the French PhVTC, i.e. the heads of the 31 centres, meets at the Agency in order to discuss the opportunity of setting up appropriate adverse event drug reaction surveys at a national level. The pharmacovigilance signals that are discussed by the PhVTC come from different sources: French spontaneous reports, European alerts, international alerts or publications.

During the period 1 January 1996–1 July 2002, the French PhVTC launched 380 such investigations involving one or several adverse event-drug combinations from which we built our sets of RSs.[21] All these 'alerts' were not launched on the basis of a quantitative analysis of the spontaneous data. The study was restricted to investigations providing sufficient details on the combinations so that they could be coded according to the ATC hierarchy,[23] 5th level, chemical subgroup, for drugs. This implied, in particular, that investigations covering a whole therapeutic class were discarded. In case of several investigations involving the same combination over the 6.5-year study period, the initial date was used as the index date. This reduced the number of investigations to 290, the latter involving 335 RSs. These 335 RSs involved 202 different drugs and 110 different adverse events. Overall, 14.6% of the RSs concerned antivirals, 6.7% psychoanaleptics, 5.4% lipid-lowerig drugs, 5.1% vaccines and 5.1% antibacterials. Concerning adverse events, 10.4% of the RSs concerned hepatitis, 4.5% haemorrhage, 3.3% diarrhoea and 3.3% pancreatitis.

Three sets of RSs were defined, corresponding to an increasing level of support from the pharmacovigilance data:

- *Set 1*: RSs associated with at least one report in the analysis following the launch of the investigation.
- *Set 2*: RSs associated with at least one report in the month following the launch of the investigation AND associated with at least three reports at the end of the study.
- *Set 3*: RSs associated with at least three reports in the month following the launch of the investigation. The latter set can be considered as reasonably supported by the data at the date of investigation.

Automatic Signal Detection Methods

Automatic signal detection methods rely on more or less complex statistical models from which disproportionality measures are derived for each combination. These disproportionality measures aim at ranking the combinations according to their potential safety issues given the information available in the pharmacovigilance database. They are calculated from various 'risk' measures defined on the basis of the underlying statistical model considered.

Details of the methods appear in the references discussed in this paragraph. Briefly, $PRR_{02.5}$ consists of calculating the lower bound of the 95% CI of the PRR.[1] The more complex Bayesian methods $IC_{02.5}$ and GPS_{05} are based on the 2.5% quantile of the *posterior* distribution of the IC probability ratio and on the 5% quantile of the

posterior distribution of the parameter, λ, governing the GPS model, respectively. GPS_{pH0} differs from GPS_{05} in that it ranks the combinations according to the posterior probability for λ to be less than 1. Finally, midRFET orders the combinations according to the mid p-values derived from univariate Fisher's exact tests. $PRR_{02.5}$, $IC_{02.5}$ and GPS_{05} were firstly described by van Puijenbroek et al.,[1] Norén et al.[4] and Szarfman et al.,[5] respectively. midRFET is presented by Ahmed et al.,[9] and GPS_{pH0} is described by Ahmed et al.[10] All methods except $PRR_{02.5}$ are also summarized in the supplementary material of Ahmed et al.[11]

When routinely used, $PRR_{02.5}$ GPS_{05} and $IC_{02.5}$ entail a detection threshold for signal generation. Here, we focused on disproportionality measures rather than the whole proposed detection strategy. This allowed us to compare the methods on the basis of an identical number of generated signals.

All methods except $IC_{0.25}$ are currently implemented in the *Ph*armaco*Vi*gilance *D*etection (PhViD) package[25] of the R software.[26] $IC_{0.25}$ was implemented in R as described by Norén et al.[4]

Analysis Plan

The analyses were performed on the whole spontaneous database (i.e. to adverse event-drug combinations associated with at least one report in the spontaneous database at the date of analysis) or, alternatively, restricted to combinations associated with at least three reports at the date of analysis.

Bibliographic Index

One of the specificities of the French pharmacovigilance system is that the pharmacovigilance experts report an imputability score[27] for each adverse drug reaction, one of whose components expresses whether the adverse event-drug combination is already known in the literature, i.e. labelled, often reported or published. It is thus possible to calculate a 'bibliographic index' for each combination corresponding to the proportion of reports in which the combination was attested to be known in the literature. In this work, the drug-event pairs for which this index was greater than 0.8 at the end of the first year of the study period, i.e. 1995 were considered as 'previously analysed' at the beginning of the study and consequently removed from any subsequent list of generated signals.

Study 1: Comparison of the Five Automatic Signals Detection

The aim of the first study was to compare the five automatic signal detection methods according to the number of RSs detected and the time to detections of the latter. The methods were sequentially applied on a monthly basis to the spontaneous reports collected between 1 January 1995 and the date of analysis. The first analysis was performed on 1 January 1996, i.e. 1 year after the collection of the first reports, and the last analysis was carried out on 1 July 2002, thus leading to a total of 79 analyses. For each analysis and for each method, the signals were ranked according to their corresponding disproportionality measures. To compare the methods on the same basis, the 100 new best-ranked signals were selected every month for each method. At the end of the study, the number of signals generated by each method was thus 7900.

For a given RS, the times to detection were calculated as the difference between the date of detection by a given method and the date of its first report in the database. If one combination was already present on 1 January 1996, the latter was used to calculate the delay instead. For a given RS, the set of methods to be compared was ordered according to the time to detection and their ranks kept, the smaller the rank the earlier the detection. The null hypothesis of no difference in times to detection between the methods was then tested by means of Friedman's non-parametric rank test performed on RSs, which were detected by all the methods.

Study 2: False Discovery Rate-Based Detection and Comparison with Traditional Detection

In this study, we evaluated whether automatic detection can detect RSs in a timely manner compared with traditional detection, i.e. detection operated by the PhVTC. For this purpose, we

mimicked the routine use of the GPS_{pH0} with a detection threshold based on the FDR. Thus, the scenario assumed that on 1 January 1996, a first analysis has been performed based on a pre-specified value of the FDR. This first analysis, which inevitably results in a considerable number of signals, can be thought of as an exploration of the data already collected before routinely using the GPS_{pH0}. The signals generated by this first analysis, as well as the combinations known according to the bibliographic index, were then considered as having been previously analysed.

Two values for the FDR threshold were studied (0.01 and 0.05). One consequence of using a pre-specified threshold of the FDR is that the number of signals to be analysed every month is no longer fixed.

We compared the time lag between the date of detection by the FDR-based GPS_{pH0} and the date of investigation by the PhVTC. The analysis was performed on RSs from sets 2 and 3 and restricted to those detected by GPS_{pH0}.

Results

Data Support of the RSs

Three hundred and thirty-five RSs were extracted from the 380 investigations launched by the PhVTC. Table I shows some characteristics of these 335 RSs in the French pharmacovigilance database. The table first indicates that 92 (335 – 243) RSs were not associated with any spontaneous report during the study period. Additionally, among the 243 RSs associated with at least one report at the end of the study period, i.e. on 1 July 2002, 48 (243 – 195) could not have been detected by any statistical methods before traditional pharmacovigilance since they were not associated with any spontaneous reports at the index date.

Before the first analysis, 1042 combinations had a bibliographic index greater than 0.8, of which 17 were RSs. These 1042 combinations were considered as having been previously evaluated. The 17 RSs known according to the bibliographic index were removed from the RS sets. The comparison was thus performed on 178 (195 – 17) RSs from set 1, 146 (163 – 17) RSs from set 2 and 101 (118 – 17) RSs from set 3 (see table I).

Study 1: Comparison of Five Automatic Signal Detection Methods

Comparison of Number of Detected RSs

Table II presents the number of RSs detected after the last analysis performed on 1 July 2002. Globally, it shows that the Bayesian methods outperform the frequentist methods, with a slight advantage in favour of GPS_{pH0}. When the analysis is done on the whole database, the latter detects 91 of the RSs of sets 1 and 2, which represent 51.1% of the RSs of set 1 and 62.3% of set 2. For set 3, both GPS-based methods detect 75 RSs, i.e. 74.3% of the 101 RSs. It is noteworthy that the RSs detected by the three Bayesian methods do not vary between set 1 and set 2, which could be due to the fact that the ranking statistics of these methods strongly penalize any signal with fewer than three reports. On the other hand, for frequentist methods, the number of detected RSs decreases between both sets, which indicates that few RSs detected by frequentist methods are associated with fewer than three reports at the end of the study. It is also worth noting that the performances of $PRR_{02.5}$ are significantly lower than those of the other frequentist

Table I. Data support in the French pharmacovigilance database of the 335 reference signals

	No. of RSs	Set
Total number of RSs	335	
With at least one report on 01/07/2002	243	
With at least three reports on 01/07/2002	181	
With at least one report at the date of investigation	195	1
With at least one report at the date of investigation and at least three reports on 01/07/2002	163	2
With at least three reports at the date of investigation	118	3

RSs = reference signals.

Table II. Number of reference signals detected by the five automatic signal detection methods

Set (n)	GPS_{pH0}	GPS_{05}	$IC_{02.5}$	midRFET	$PRR_{02.5}$
Analysis on the whole database					
1 (178)	91	90	90	87	68
2 (146)	91	90	90	84	66
3 (101)	75	75	73	69	55
Analysis restricted to adverse event-drug combinations with at least three reports					
2 (146)	91	90	90	88	87
3 (101)	75	75	73	71	70

GPS_{pH0} = posterior probability of the null hypothesis H_0 calculated from the Gamma Poisson Shrinker; **GPS_{05}** = 5% quantile of the Gamma Poisson Shrinker; **$IC_{02.5}$** = 2.5% quantile of the Information Component; **midRFET** = mid p-values calculated from the Fisher's Exact Test; **$PRR_{02.5}$** = lower bound of the 95% CI of proportional reporting ratio.

method, i.e. midRFET. With the latter being based on the Fisher's exact test, it is more adequate than $PRR_{02.5}$ when applied to a database containing a large amount of combinations with small counts. When the analysis is restricted to drug-event pairs with at least three reports, the performances of the frequentist methods improve but remain slightly under those of the Bayesian ones.

For the analysis performed on the whole database, it can be deduced from tables II and III that all RSs detected by $PRR_{02.5}$ (66 for set 2 and 55 for set 3) are detected by GPS_{pH0} (91 for set 2 and 75 for set 3), while for set 1 (counts not reported in table III) two RSs detected by $PRR_{02.5}$ are not detected by GPS_{pH0}. For the restricted analysis, there are three and one signals detected by $PRR_{02.5}$ that are not detected by GPS_{pH0} for set 2 and set 3, respectively. Overall, GPS_{pH0} identified more signals than the other methods, in particular versus $PRR_{02.5}$, with a large extent of overlap between the methods.

Comparison of Times to Detection

The methods were then compared with regard to their time to detection of the RSs. Results in table III show the average detection rank for each method, as well as the p-values resulting from Friedman's test. The analysis was not performed on the RSs in set 1 since none of the signals with fewer than three reports was detected by any of the Bayesian methods. It was carried out either

Table III. Comparison of time to detection. For each analysis, the average rank of detection (varying between 1 and 5) is calculated on the common detected reference signals. The hypothesis of no difference in time to detection is tested by means of Friedman's test

Set	Number of shared RSs	Average rank of detection					p-Values
		GPS_{pH0}	GPS_{05}	$IC_{02.5}$	midRFET	$PRR_{02.5}$	
Analysis on the whole database							
2	66	2.53	2.63	2.79	3.07	3.99	6.46e–13
	90	1.42	1.58	NA	NA	NA	0.010
3	55	2.43	2.64	2.76	3.10	4.07	1.18e–12
	75	1.39	1.61	NA	NA	NA	1.07e-3
Analysis restricted to adverse event-drug combinations with at least three reports							
2	84	2.81	2.84	2.95	3.03	3.37	5.61e–3
	90	1.46	1.54	NA	NA	NA	0.103
3	69	2.65	2.75	3.00	3.09	3.51	5.37e–5
	75	1.43	1.57	NA	NA	NA	0.033

GPS_{pH0} = posterior probability of the null hypothesis H_0 calculated from the Gamma Poisson Shrinker; **GPS_{05}** = 5% quantile of the Gamma Poisson Shrinker; **$IC_{02.5}$** = 2.5% quantile of the Information Component; **midRFET** = mid p-values calculated from the Fisher's Exact Test; **NA** = not applicable; **$PRR_{02.5}$** = lower bound of the 95% CI of proportional reporting ratio; **RSs** = reference signals.

on the whole database or restricted to combinations with at least three reports. Compared with other methods, GPS_{pH0} had the best average time to detection and the p-values associated with the tests are in favour of rejection of the null hypothesis of no difference in times to detection between the methods. The tests applied to both detection rules derived from the GPS model were in favour of GPS_{pH0} when the sequential analysis was done on the whole pharmacovigilance database and for the RSs in set 3 in the restricted analysis.

Study 2: FDR-Based Detection and Comparison with Traditional Detection

The comparison between GPS_{pH0} and traditional pharmacovigilance was based on the 146 RSs from set 2 and the 101 RSs from set 3. The analysis was restricted to combinations associated with at least three reports.

Results regarding the number of detected RSs are summarized in table IV. Compared with the first study, the analysis based on both FDR thresholds leads to an increase in the number of signals to be analysed each month, especially when the FDR is set at 0.05. The percentage of detected RSs varied between 71.9% and 87.1%.

Figure 1 presents the cumulative proportion of detected RSs according to the time lag (in months) of detection. A negative delay indicates early detection by GPS_{pH0}. For instance, it shows that for the RSs in set 2, more than 60% of the detected signals by the GPS_{pH0} were detected before the traditional detection. When the analysis is performed on the RSs in set 3, more than 80% of the RSs detected by GPS_{pH0} were detected before the traditional pharmacovigilance screening. For the RSs in set 2, the median delays are equal to −3 months for both FDR-based thresholds. When the analysis was based on the RSs in set 3, the median delays are equal to −7 and −9 months for FDR thresholds of 0.01 and 0.05, respectively.

Finally, table V details the comparison of the time to detection according to the number of spontaneous reports at the end of the study for the RSs in set 3 and an FDR threshold of 0.05. It illustrates first that the signals associated with more than 100 reports were all detected. For the latter, we also found that they were all detected earlier with GPS_{pH0}. More generally, table V shows that GPS_{pH0} performances in terms of early detection and proportion of detected RSs increase with the number of reports.

Discussion

Several studies have demonstrated the value of automatic signal detection methods based either on simulations or retrospective analyses. They have also shown that evaluating these methods is intimately related to defining the RSs. The RSs used in this work correspond to combinations that were considered relevant by a group of pharmacovigilance experts at a particular moment. They make it possible to really compare traditional detection to automatic signal detection. However, this definition of RSs presents some drawbacks. First of all, it is in a sense unfair for the automatic detection methods since some

Table IV. Summary of detection of False Discovery Rate-based GPS_{pH0}

FDR thresholds	0.01	0.05
Number of signals for the first analysis[a]	1567	2259
Number of signals on 01/07/2002 (exclusion of the signals of the first analysis)	9543	16 706
Average number of signals to be analysed each month	124	216
Number of RSs detected in the first analysis	22	26
Number of RSs detected at the end of the study	Set 2: 105 (71.9%) Set 3: 79 (78.2%)	Set 2: 122 (83.6%) Set 3: 88 (87.1%)

a The signals include the 1042 combinations known according to the bibliographic index.

FDR = False Disovery Rate; **GPS_{pH0}** = posterior probability of the null hypothesis H_0 calculated from the Gamma Poisson Shrinker; **RSs** = reference signals.

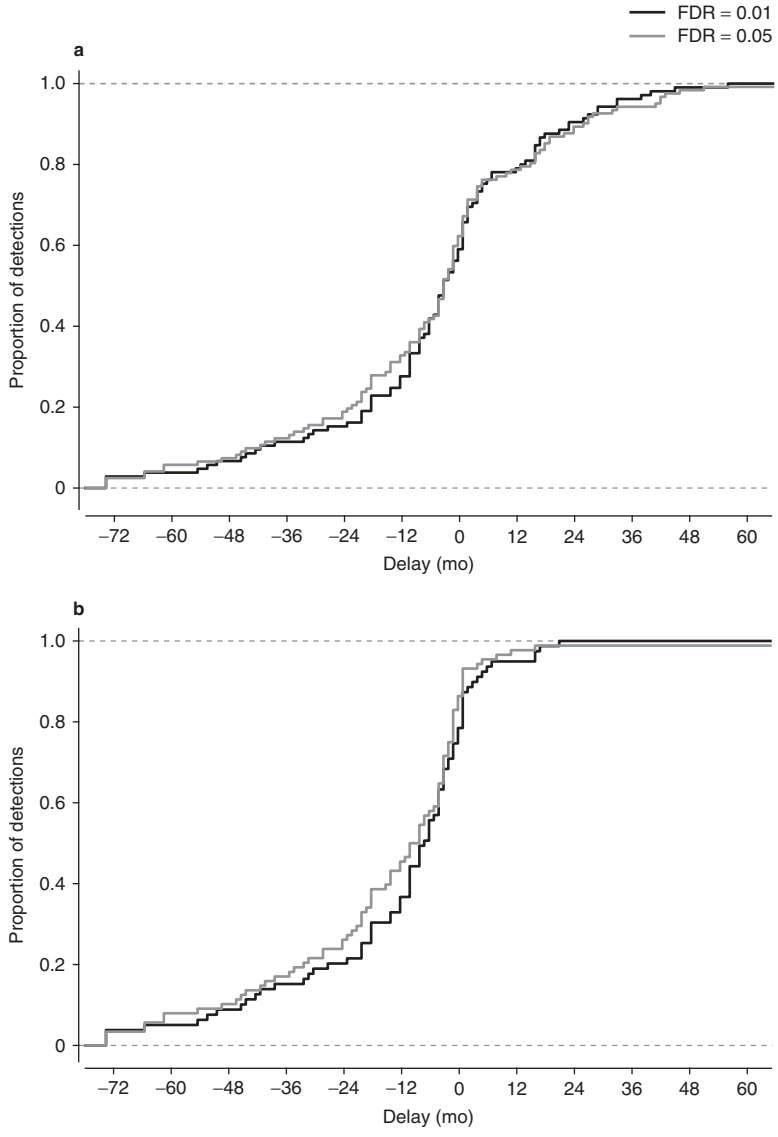

Fig. 1. Cumulative proportion of detected reference signals from (**a**) set 2 and (**b**) set 3 according to time lag between date of detection of GPS_{pH0} and launch of investigation by the French Pharmacovigilance Technical Committee. A negative delay indicates early detection by the False Discovery Rate-based GPS_{pH0}. The proportion is calculated on reference signals detected by GPS_{pH0} at the end of the study. **FDR** = False Discovery Rate; **GPS_{pH0}** = posterior probability of the null hypothesis H_0 calculated from the Gamma Poisson Shrinker.

alerts are launched by the French PhVTC on the basis of information external to the spontaneous reporting database, e.g. European alerts or published reports. As a consequence, some RSs are based on very little data and cannot be detected by any automatic signal detection tool. Secondly, this definition implicitly assumes that the automatic signal detection tools aim to detect what pharmacovigilance experts detect. However, the combinations targeted by these methods do not

correspond only to those detected by pharmacovigilance experts, as one of their main purposes is to draw attention to unsuspected associations. This aspect of automatic signal detection seems to be impossible to evaluate by empirical studies so we have to assume that if the methods make it possible to detect relevant signals for pharmacovigilance experts, they will also be efficient in their role of hypothesis generators.

Based on this definition of RSs, this study indicates that the Bayesian methods globally perform better than the frequentist methods not only in terms of number of detected signals but also in terms of time to detection. It also shows that GPS_{pH0} slightly outperforms the other methods. Interestingly, these conclusions are consistent with the simulation results of Ahmed et al.[11] Additionally, as for previous studies, the present work confirms that automatic signal methods are useful for detecting potential adverse drug reactions that have some supporting data. Frequentist methods seem to be able to detect a few RSs with fewer than three reports. For instance, among the 32 RSs in set 1 associated with fewer than three reports at the end of the study, $PRR_{02.5}$ was able to detect two signals. Additionally, for these two signals, the comparison of the date of detection to the date of investigation was in favour of $PRR_{02.5}$. However, the detection of these few signals has to be counterbalanced with much lower overall performances than the Bayesian methods in terms of number of RSs detected and time to detection. The detection of associations that are based on very little data is not satisfactorily fulfilled by automatic signal detection methods. In our opinion, however, this task can be efficiently achieved by pharmacovigilance experts.

Several limitations exist in the present study. First, we did not exploit supplementary information such as the age and sex of patients. It would also be beneficial to implement some tools to pre-process the French data in order to decrease the number of spurious signals. Finally, we restricted the evaluation of the methods on the basis of precise combinations (ATC5 level for the drugs), which led us to discard a non-negligible number of therapeutic class investigations.

Few studies have considered retrospective analyses with a sequential perspective on a large set of RSs.[19,20] In addition to the differences concerning the definition of RSs discussed above, the present study differs from these previous studies in several aspects. First, we evaluated the methods used on the three largest pharmacovigilance databases as well as two other recent methods. The study by Alvarez et al.[20] was restricted to $PRR_{02.5}$ (applied to signals associated with at least three reports), which was shown in our work to give, overall, the worst results among the five methods. Alvarez et al.[20] incidentally indicate their willingness to investigate more complex methods in the future. In the work by Hochberg and Hauben,[19] three methods are compared but based on their current detection thresholds. We considered it fairer to compare the different methods based on a realistic and identical number of signals to be evaluated every

Table V. Median and average delays[a] (in months) between the date of detection by GPS_{pH0} (False Discovery Rate = 0.05) and the date of investigation launched by traditional detection (index date) for reference signals in set 3 stratified according to their corresponding number of reports in the pharmacovigilance database at the end of the study (1 July 2002)

Number of reports on 1 July 2002	Number of RSs in set 3	Number of detected RSs by GPS_{pH0} [n (%)]	Median delay (months)	Average delay (months)
≥3	101	88 (87.1)	−9	−15.8
≥10	71	63 (88.7)	−10	−17.0
≥20	47	43 (91.5)	−11	−17.5
≥50	25	23 (92.0)	−12	−23.7
≥100	9	9 (100.0)	−12	−19.8

a Delays are calculated for RSs detected by GPS_{pH0}. A negative delay indicates early detection by GPS_{pH0}. The average and median delay are calculated based on RSs detected by GPS_{pH0}.

GPS_{pH0} = posterior probability of the null hypothesis H_0 calculated from the Gamma Poisson Shrinker; **RSs** = reference signals.

month. By doing so, we aimed to identify the best disproportionality measure rather than the best global detection strategy, which depends heavily on the choice of a detection threshold, the latter being guided eventually by the amount of time that can be spent analysing the signals further.

We also show the impact of several detection rules, i.e. based on the analysis of 100 signals per month or based on two different thresholds of the FDR for GPS_{pH0}. As pointed out by Alvarez et al.,[20] further analysis of signals generated by detection methods is a burdensome task and eventually governs the choice of a detection rule. The number of signals to be evaluated every month can be modulated either by fixing it, i.e. 100 as in the first study, or by using different thresholds for the ranking statistic. An interesting feature of FDR-based methods is that they make it possible to estimate a statistical criterion accounting for the numerous comparisons that are performed. This estimate can be used either as a threshold for detection or as an indicator of the reliability of the signals generated by the automatic tool.

Conclusions

Our results confirm the practical potential of automatic signal detection methods. They are certainly not designed to replace pharmacovigilance expertise, as a large proportion of RSs evaluated in this study could not possibly be detected in a timely manner on the sole basis of data mining algorithms. On the other hand, we clearly show that as soon as there is reasonable support for the data, they are powerful tools to efficiently explore large spontaneous reporting system databases and detect relevant signals quickly compared with traditional pharmacovigilance.

Acknowledgements

This work was supported in part by IRESP (Institut de Recherche en Santé Publique) grant A06076LS. The authors have no conflicts of interest that are directly relevant to the content of this study.

References

1. van Puijenbroek EP, Bate A, Leufkens HGM, et al. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf 2002 Feb; 11 (1): 3-10
2. Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf 2001 Nov; 10 (6): 483-6
3. Bate A, Lindquist M, Edwards IR, et al. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol 1998 Jun; 54 (4): 315-21
4. Norén GN, Bate A, Orre R, et al. Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. Stat Med 2006 Nov 15; 25 (21): 3740-57
5. Szarfman A, Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database. Drug Saf 2002; 25 (6): 381-92
6. DuMouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. Am Stat 1999 Aug; 53 (3): 177-90
7. DuMouchel W, Pregibon D. Empirical bayes screening for multi-item associations. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco (CA): ACM, 2001: 67-76 [online].Available from URL: http://portal.acm.org/citation.cfm?id=502526 [Accessed 2010 Mar 26]
8. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 1995; 57 (1): 289-300
9. Ahmed I, Dalmasso C, Haramburu F, et al. False discovery rate estimation for frequentist pharmacovigilance signal detection methods. Biometrics 2010 Mar; 66 (1): 301-9
10. Ahmed I, Haramburu F, Fourrier-Réglat A, et al. Bayesian pharmacovigilance signal detection methods revisited in a multiple comparison setting. Stat Med 2009 Jun 15; 28 (13): 1774-92
11. Ahmed I, Thiessard F, Miremont-Salamé G, et al. Pharmacovigilance data mining with methods based on false discovery rates: a comparative simulation study. Clin Pharmacol Ther 2010 Oct; 88 (4): 492-8
12. Roux E, Thiessard F, Fourrier A, et al. Evaluation of statistical association measures for the automatic signal generation in pharmacovigilance. IEEE Trans Inf Technol Biomed 2005 Dec; 9 (4): 518-27
13. Matsushita Y, Kuroda Y, Niwa S, et al. Criteria revision and performance comparison of three methods of signal detection applied to the spontaneous reporting database of a pharmaceutical manufacturer. Drug Saf 2007; 30 (8): 715-26
14. Almenoff JS, LaCroix KK, Yuen NA, et al. Comparative performance of two quantitative safety signalling methods: implications for use in a pharmacovigilance department. Drug Saf 2006; 29 (10): 875-87
15. Lehman HP, Chen J, Gould AL, et al. An evaluation of computer-aided disproportionality analysis for post-marketing signal detection. Clin Pharmacol Ther 2007 Aug; 82 (2): 173-80

16. Lindquist M, Ståhl M, Bate A, et al. A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database. Drug Saf 2000 Dec; 23 (6): 533-42
17. Hauben M, Madigan D, Gerrits CM, et al. The role of data mining in pharmacovigilance. Expert Opin Drug Saf 2005 Sep; 4 (5): 929-48
18. Hochberg AM, Hauben M, Pearson RK, et al. An evaluation of three signal-detection algorithms using a highly inclusive reference event database. Drug Saf 2009; 32 (6): 509-25
19. Hochberg AM, Hauben M. Time-to-signal comparison for drug safety data-mining algorithms vs. traditional signaling criteria. Clin Pharmacol Ther 2009 Jun; 85 (6): 600-6
20. Alvarez Y, Hidalgo A, Maignen F, et al. Validation of statistical signal detection procedures in eudravigilance post-authorization data: a retrospective evaluation of the potential for earlier signalling. Drug Saf 2010 Jun 1; 33 (6): 475-87
21. Thiessard F. Détection des effets indésirables des médicaments par un système de génération automatisée du signal adapté à la base nationale de pharmacovigilance [in French; thesis]. Bordeaux: Université Victor Segalen Bordeaux 2, 2004
22. Thiessard F, Roux E, Miremont-Salamé G, et al. Trends in spontaneous adverse drug reaction reports to the French pharmacovigilance system (1986-2001). Drug Saf 2005; 28 (8): 731-40
23. Miller GC, Britt H. A new drug classification for computer systems: the ATC extension code. Int J Biomed Comput 1995 Oct; 40 (2): 121-4
24. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). Drug Saf 1999 Feb; 20 (2): 109-17
25. Ahmed I, Poncet A. PhViD: an R package for pharmacovigilance signal detection [online]. Available from URL: http://cran.r-project.org/web/packages/PhViD/index.html [Accessed 2012 Mar 29]
26. R Development Core Team. R: a language and environment for statistical computing [online]. Vienna, 2011. Available from URL: http://www.R-project.org [Accessed 2012 Mar 29]
27. Bégaud B, Evreux JC, Jouglard J, et al. Imputation of the unexpected or toxic effects of drugs: actualization of the method used in France. Therapie 1985 Apr; 40 (2): 111-8

Correspondence: Dr *Ismaïl Ahmed*, CESP Centre for Research in Epidemiology and Population Health, U1018, Biostatistics, 16, Av P.V. Couturier, 94807 Villejuif, France.
E-mail: ismail.ahmed@inserm.fr

