Evaluation of Statistical Association Measures for the Automatic Signal Generation in Pharmacovigilance
Emmanuel Roux, Frantz Thiessard, Annie Fourrier, Bernard Bégaud, Pascale Tubert-Bitter

To cite this version:

HAL Id: hal-00396036
https://hal.archives-ouvertes.fr/hal-00396036
Submitted on 16 Jun 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Evaluation of Statistical Association Measures for the Automatic Signal Generation in Pharmacovigilance

Emmanuel Roux, Frantz Thiessard, Annie Fourrier, Bernard Bégaud, and Pascale Tubert-Bitter

Abstract—Pharmacovigilance aims at detecting the adverse effects of marketed drugs. It is generally based on the spontaneous reporting of events thought to be the adverse effects of drugs. Spontaneous Reporting Systems (SRSs) supply huge databases that pharmacovigilance experts cannot exhaustively exploit without data mining tools. Data mining methods; i.e., statistical association measures in conjunction with signal generation criteria, have been proposed in the literature but there is no consensus regarding their applicability and efficiency, especially since such methods are difficult to evaluate on the basis of actual data. The objective of this paper is to evaluate association measures on simulated datasets obtained with SRS modeling. We compared association measures using the percentage of false positive signals among a given number of the most highly ranked drug–event combinations according to the values of the association measures. By considering 150 drugs and 100 adverse events, these percentages of false positives, among the 500 most highly ranked drug-event couples, vary from 1.1% to 53.4% (averages over 1000 simulated datasets). As the measures led to very different results, we could identify which measures appeared to be the most relevant for pharmacovigilance.

Index Terms—Adverse drug reaction reporting systems, association measures, computer simulation, information systems, validation studies.

I. INTRODUCTION

C linical trials are efficient for identifying the most frequent adverse effects of a drug prior to marketing. However, for obvious reasons, the effects of rare occurrences cannot be detected. Such effects can be specific to a population subgroup and/or can have a latency longer than the trial duration. The identification of such effects is the scope of pharmacovigilance, whose role includes the post-marketing surveillance of adverse effects based on the spontaneous reporting by the medical community of adverse events suspected to be related to a medication. The main drawbacks of this spontaneous reporting system (SRS) are considerable underreporting, to an unknown extent, and the fact that a report does not prove the causal relationship between drug(s) and event(s). Moreover, SRS supplies huge databases with a continuous flow. For example, in 1997, 35,000 new reports were added quarterly to the World Health Organization (WHO) database [1]. In 1999, DuMouchel mentioned that the Food and Drug Administration (FDA) database contained 1.2 million different reports [2]. On 1 January 2000, The Netherlands Pharmacovigilance Foundation LAREB contained 26,555 reports concerning 17,330 different drug–event combinations [3]. At the end of 2001, the French pharmacovigilance database contained about 200,000 reports referring to about 185,000 different drug–event couples [4]. Such massive databases preclude human-based exploration, so several automated signal generation methods have recently been proposed to assist pharmacovigilance experts.

A good evaluation of the suspicious character of a drug–event couple would be to determine to what extent the observed number of reports referring to this couple exceeds the expected number of reports, assuming the independence between exposure to the drug and the adverse event. However, the background incidence of the event in the whole population, the number of patients exposed to the drug, and the extent of under-reporting are unknown. This prevents reliable computation of the expected number of reports for drug–event couples. One solution proposed is to use Data Mining (DM) methods that only exploit the intrinsic information of the database in order to estimate, for a given couple, the expected number of reports by means of the data related to all the other drugs and events. Pharmacovigilance experts are able to identify unusually frequent drug–event reports and serious events, but numerous adverse reactions are difficult to identify. Those reactions are not obviously frequent or even rare, appear with a substantial latency period, are not serious enough to draw the attention of the experts, etc. So DM methods are supposed to draw experts’ attention to more “surprising” drug–event couples. Such methods are not supposed to replace experts, but act as hypotheses generators. These hypotheses are destined to be validated or invalidated by the experts after deeper investigations.

Several methods have been proposed, from very simple and intuitive ones using disproportionality measures to Bayesian ones, demanding more statistical and computational skills. However, since no consensus exists concerning the DM method(s) to be used, routine application of such methods in pharmacovigilance is still limited. In fact, the unknown features of the data, i.e., the events background incidence, the number of patients exposed to the drug, the extent of under-reporting and the true status of the drug-event relationship, not only prevent to compute a reliable expected number of reports, but also prevent to quantitatively, objectively, and absolutely evaluate DM methods. Moreover, the suitability of these methods for pharmacovigilance is questioned [5].

Waller et al. [6] recalled that “A judgment on the validity and utility of these measures should be based on comparison
of their sensitivity, specificity and predictive values in signal
detection from a real dataset" but also précised that "such (real)
data do not yet exist." Nevertheless, some authors attempted to
evaluate automatic signal generation methods on real data. The
main contributions have been published by Van Puijenbroek et
al. [3] and Gould [7]. Van Puijenbroek et al. compare different
methods in terms of sensitivity, specificity, positive, and negative
predictive values by considering the results of one of these
methods as the reference. This permits a relative evaluation of
the methods, but does not describe their absolute performances.

In this paper, we evaluate ten DM methods on simulated
datasets. To our knowledge, these methods are all the proposed
ones in the context of pharmacovigilance. Simulated data are
generated by the SRS model proposed in [8]. In fact, by knowing
all the model parameters, and especially the status of the drug–
event relationships, it is possible to label generated signals as
"true" or "false."

However, to be accepted and used by pharmacovigilance ex-
erts, an automated signal generation method must not only be
reliable, but also well adapted to specific practical issues. The
main drawback in most of the existing methods is the excessive
number of signals they generate, thus preventing the assessment
of all of them by pharmacovigilance experts. For this reason, a
suitable method would be one that either generates an acceptable
number of "true" signals, or which correctly ranks drug–event
couples on a continuum from more to less suspicious. In fact,
by analyzing a limited number of the most highly ranked com-
binations depending on the human, material and time resources
available, the risk of uselessly analyzing false signals would be
minimized if the ranking is pertinent according to the relative
risks of the events.

In this paper, we first briefly describe the data generation
process and the simulated data. Then, we present the DM meth-
ods and discuss their evaluation criteria in view of the above-
mentioned issues. We then present our results and discuss them.

II. METHODS

A. Simulated Data

Data to which we applied the signal generation methods were
generated by the spontaneous reporting system modeling pro-
posed in [8], where the reporting process is viewed as a Poisson
process. For a given drug–event couple and during a given pe-
riod, the expected number of reports is defined as follows:

\[ \delta = e \cdot i \cdot RR \cdot pr_e \]  

where \( e \), \( i \), and \( RR \) are the drug exposure frequency, the
event background incidence, and the event relative risk, re-
spectively. The event relative risk is the rate of the proba-
probability to observe the event given the exposure to the drug.
over the probability to observe the event without the expo-
sure to the drug. \( pr_e \) is the reporting probability. This report-
ing probability, which is quantitatively unknown, is derived
in the model from the qualitative knowledge reported in the
literature and expressed by experts. The knowledge is rep-
resented and exploited by means of a fuzzy characterization
of variables and a set of fuzzy rules. For a given drug–event
combination, the reporting probability depends on the delay
since the drug launch, the number of reports in the past and the
seriousness of the event.

In the present study, 10 years of spontaneous reporting sys-
tem were simulated, with a generation period of six months. We
arbitrarily considered 150 drugs and 100 adverse events. The
maximal exposures to the drugs over the 10-year period, \( e \), were
two million, 300 000 and 30 000, each value being applied to
one third of the drugs. For each maximal exposure value, five
drugs were launched each year during the 10 year period, lead-
ing to 10 different delays since launch, denoted \( d \). Background
incidences of adverse events, \( i \), were \( 1/50 \) 000 and \( 1/10 \) 000,
each being associated to one half of the events. For each value
of background incidence, half of the events were serious and
the remaining ones were mild. Ten percent of the drug–event
couples were associated with a relative risk, \( RR \), ranging from
1.2 to 10, and for 90% of couples, \( RR \) was equal to 1. This
distribution was imposed on each data subset having the same
value for \( e, i, \) seriousness, and \( d \).

One thousand datasets were generated. On average, 10502
(Standard Deviation (SD = 35) drug–event combinations were
reported at least once. Two thousand (SD = 37) drug–event
couples were reported only once, 1181(SD = 29) twice, 770(SD = 25) three times and 6551(SD = 23) at least four
times. These numbers correspond to 19.0%, 11.3%, 7.3% and
62.4% of the reported couples, respectively. The average
maximal number of reports for a drug–event couple was
537(SD = 19). The average number of “true associations” i.e.,
the number of couples whose relative risk exceeds one, was
1182(SD = 10), thus corresponding to 11.3% of the reported
couples. Ideally, a signal generation method should detect all
11.3% of the associations.

Fig. 1 presents the relative cumulated number of drug–event
combinations (averaged over the 1000 simulated datasets)
according to the number of reports and as a function of the
model parameters.
TABLE I  
2 × 2 CONTINGENCY TABLE

<table>
<thead>
<tr>
<th></th>
<th>Event</th>
<th>Other events</th>
<th>All events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>a</td>
<td>b</td>
<td>a+b</td>
</tr>
<tr>
<td>Other drugs</td>
<td>c</td>
<td>d</td>
<td>c+d</td>
</tr>
<tr>
<td>All drugs</td>
<td>a+c</td>
<td>b+d</td>
<td>N</td>
</tr>
</tbody>
</table>

B. Signal Generation Methods

Ten signal generation methods were applied to the simulated data. They are, to our knowledge, all the methods proposed in the context of pharmacovigilance. They are described in the Appendix and all exploit the 2 × 2 contingency table as in Table I. Although the data generation process was sequential in order to take into account the time in the reporting process modeling [8], the signal generation methods were applied at the end of the 10 year period.

In addition to the observed numbers of reports, some methods exploit the expected number of reports according to the information present in the database. For a given drug–event couple, this expected number of reports, \( E \), is the theoretical number of reports for independent rows and columns in the dataset. It is computed with the marginal numbers and the total amount of reports, \( N \), in the database. By using notations of Table I, \( E \) is obtained with the following equation:

\[
E = \frac{(a + b) \cdot (a + c)}{N}.
\]  

The automatic signal generation methods we compared are based on the proportional reporting ratio (PRR) [9], the reporting odds ratio (ROR) [3], [10], [11], Yule’s \( Q \) [3], the sequential probability ratio test (SPRT) [12], [13], the tests using Poisson and Chi-square distributions (denoted Poisson and \( \chi^2 \), respectively) [3], the information component (IC) [1], [7], [14], the empirical bayes arithmetic mean (EBAM) [2], [15] and an alternative method called empirical bayes probability (EBp) derived from an intuitive interpretation of the mixture model of DuMouchel [2], [15].

C. Evaluation Criteria

The ten association measures presented in the Appendix were applied to 1000 simulated datasets, thus allowing the computation of averages and variances of the evaluation indexes. In the following, the same names are used for the signal generation methods and the corresponding association measures.

In Section III, we first present and discuss the receiver operating characteristic (ROC) curves of the association measures, computed with 200 simulated datasets. However, as shown in Section III and in Table II, the signal generation methods, i.e., association measures in conjunction with generation criteria, provided too many signals, making them inapplicable in the routine, as such. In fact, it would be impossible for pharmacovigilance experts to assess all the generated hypotheses. For this reason, it seems more practical to rank the drug–event combinations in a decreasing order according to the association measure values for EBAM, EBp, IC, \( \chi^2 \), PRR, ROR, Yule’s \( Q \), and SPRT\(_2\), and in an increasing order according to the probabilities for Poisson. Then investigations can be performed on an acceptable number of the most highly ranked hypotheses. This method was proposed by DuMouchel [2], [15]. In such a context, evaluating signal generation methods by means of sensitivity and specificity does not make sense. In fact, experts can choose to investigate a number of combinations that is quite lower than that of the “true associations” in the dataset. Sensitivity and specificity, which aim at evaluating the methods’ ability to detect all the “true” and “false” associations, respectively, are not appropriate.

Therefore, in Section III, we evaluate the signal generation methods by means of the percentage of false positive signals among a predefined number of the most highly ranked drug–event couples. This index reflects the cost of the useless analysis performed by pharmacovigilance experts and is of major practical interest. We studied these proportions for different numbers of the most highly ranked couples: 10, 20, 50, 100, 200, and 500. The results presented are the averages of the percentages of the false positive signals over the 1000 simulated datasets.

III. RESULTS

Fig. 2 presents the ROC curves for all the association measures except ROR, which presents similar results as PRR. For each curve, the point that corresponds to the threshold proposed in literature is represented by a circle. \( \chi^2 \) gives the worst performances. All the other measures provide comparable results. EBAM performances appear to be highly variable when the sensitivity increases. However, in the present paper, by studying relatively low numbers of the most highly ranked drug–event couples, we consider very restrictive signal generation criteria that do not lead to significant variability.

It appears that for all the measures, sensitivity could be greatly improved while keeping an acceptable specificity by using less restrictive criteria.

Table II presents the number of potentially suspicious combinations for the generation criteria presented in the Appendix. These numbers are to be related to the number of reported drug–event combinations in the simulated datasets (10 502, SD = 35). In the “real” pharmacovigilance databases, this number of combinations is much larger, thus leading to a much larger number of signals.

Increasing sensitivity would also increase the number of generated signals and, consequently, worsen this problem. This justifies the practical viewpoint adopted in the present article by studying a manageable number of the most highly ranked drug–event couples according to measures values.

Table III presents, on average, over the 1000 simulated datasets, the numbers of reports, the relative risks, and the values of the measures associated with the 10 most highly ranked drug–event combinations for each measure. Obviously, EBAM, SPRT\(_2\), \( \text{IC}_{\text{Bate}} \) (\( \text{IC}_{\text{Bate}} \) is for IC with Bate’s priors), and \( \chi^2 \) rank the couples more correctly than \( \text{IC}_{\text{Gould}} \) (IC with Gould’s priors), PRR, and ROR, the relative risk being high, especially for EBAM. On the other hand, PRR, ROR, and \( \text{IC}_{\text{Gould}} \) retrieve couples with a low relative risk. Moreover, these three measures only reveal couples with a low number of reports. Note that...
TABLE II
AVERAGE NUMBER OF SIGNALS, SENSITIVITIES, AND SPECIFICITIES FOR THE GENERATION CRITERIA PROPOSED IN THE LITERATURE. THE NUMBERS OF REPORTED DRUG-EVENT COMBINATIONS AND OF “TRUE DRUG-EVENT ASSOCIATIONS” ARE 10,502 (SD = 35) AND 1182 (SD = 10), RESPECTIVELY

<table>
<thead>
<tr>
<th>Measures</th>
<th>Generation criteria</th>
<th>Number of signals Mean (SD)</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBAM</td>
<td>$L_{05}(EBAM)&gt;1$</td>
<td>386 (8)</td>
<td>0.33</td>
<td>1.00</td>
</tr>
<tr>
<td>EBp</td>
<td>EBp &gt; 0.5</td>
<td>611 (15)</td>
<td>0.49</td>
<td>1.00</td>
</tr>
<tr>
<td>SPRT2</td>
<td>SPRT2 &gt; 2.94</td>
<td>459 (8)</td>
<td>0.39</td>
<td>1.00</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>$Pr(\chi^2 &gt; \chi^2) &lt; 0.05$</td>
<td>1179 (22)</td>
<td>0.54</td>
<td>0.94</td>
</tr>
<tr>
<td>Poisson</td>
<td>$Pr(Pois &gt; a) &lt; 0.05$</td>
<td>762 (14)</td>
<td>0.58</td>
<td>0.99</td>
</tr>
<tr>
<td>IC$_{Gould}$ (*)</td>
<td>$L_{05}(IC)&gt;0$</td>
<td>1181 (20)</td>
<td>0.65</td>
<td>0.96</td>
</tr>
<tr>
<td>IC$_{Bate}$ (*)</td>
<td>$L_{05}(IC)&gt;0$</td>
<td>561 (10)</td>
<td>0.46</td>
<td>1.00</td>
</tr>
<tr>
<td>PRR</td>
<td>PRR &gt; 2 and $\chi^2 &gt; 4$ and $a \geq 3$</td>
<td>570 (12)</td>
<td>0.45</td>
<td>0.99</td>
</tr>
<tr>
<td>ROR</td>
<td>$L_{05}(ROR)&gt;1$</td>
<td>876 (16)</td>
<td>0.58</td>
<td>0.98</td>
</tr>
<tr>
<td>Yule’s Q</td>
<td>$L_{05}(Q)&gt;0$</td>
<td>1058 (19)</td>
<td>0.63</td>
<td>0.97</td>
</tr>
</tbody>
</table>

(*) IC$_{Gould}$ and IC$_{Bate}$ are for IC with Gould’s and Bate’s priors, respectively (see Appendix)

Fig. 2. ROC curves for the association measures presented in the Appendix, except ROR for which the ROC curve is similar to PRR. All the ROC curves are represented in gray in all the sub-figures, and the curve highlighted corresponds to the title of the sub-figure. Probability thresholds for curves construction are $\{0.025, 0.05, 0.1, 0.2, ..., 0.9, 0.95, 0.975\}$. For each measure, the point that corresponds to the threshold proposed in literature (see Appendix) is represented by a circle. For each threshold, the mean plus/minus the standard deviation, on 200 simulated datasets, is represented.

Table V shows results for all the drug–event associations reported at least once. Note the great differences between EBAM, EBp, SPRT2, $\chi^2$, Poisson, and IC$_{Bate}$, on one hand, and IC$_{Gould}$, PRR, and ROR, on the other. The most obvious difference concerns the percentage of false positive signals. Another difference is the influence, on this percentage, of the number of combinations taken into account. For EBAM, EBp, SPRT2, $\chi^2$, Poisson and IC$_{Bate}$, the proportion of false positive signals increases with the number of couples, while the proportion for the other measures decreases.

As shown in Table VI, measures performances are comparable for associations with high report numbers ($a \geq 5$ or $e = 3$).
TABLE III
AVERAGE OF THE NUMBER OF REPORTS, $a$, THE RELATIVE RISK, RR AND THE VALUE OF THE ASSOCIATION MEASURE, $V_{AL}$, FOR THE 10 MOST HIGHER RANKED COMBINATIONS AND FOR EACH MEASURE

<table>
<thead>
<tr>
<th>Measure</th>
<th>$a$</th>
<th>RR Val.</th>
<th>$a$</th>
<th>RR Val.</th>
<th>$a$</th>
<th>RR Val.</th>
<th>$a$</th>
<th>RR Val.</th>
<th>$a$</th>
<th>RR Val.</th>
<th>$a$</th>
<th>RR Val.</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBAM</td>
<td>308</td>
<td>9.87</td>
<td>651</td>
<td>10.00</td>
<td>280</td>
<td>8.81</td>
<td>519</td>
<td>9.97</td>
<td>2402</td>
<td>8.91</td>
<td>257</td>
<td>132.24</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>270</td>
<td>9.74</td>
<td>602</td>
<td>9.88</td>
<td>245</td>
<td>69.78</td>
<td>473</td>
<td>9.78</td>
<td>1903</td>
<td>9.34</td>
<td>257</td>
<td>54.69</td>
</tr>
<tr>
<td>$IC_{Gould}$</td>
<td>244</td>
<td>9.67</td>
<td>590</td>
<td>9.47</td>
<td>228</td>
<td>4.94</td>
<td>446</td>
<td>9.48</td>
<td>1719</td>
<td>2.3</td>
<td>257</td>
<td>44.01</td>
</tr>
<tr>
<td>$IC_{Bate}$</td>
<td>228</td>
<td>9.63</td>
<td>580</td>
<td>8.84</td>
<td>215</td>
<td>9.91</td>
<td>426</td>
<td>9.22</td>
<td>1574</td>
<td>9.98</td>
<td>257</td>
<td>37.69</td>
</tr>
<tr>
<td>$PRR$</td>
<td>216</td>
<td>9.60</td>
<td>572</td>
<td>8.73</td>
<td>205</td>
<td>9.08</td>
<td>410</td>
<td>8.92</td>
<td>1452</td>
<td>7.5</td>
<td>257</td>
<td>33.36</td>
</tr>
<tr>
<td>ROR</td>
<td>202</td>
<td>9.59</td>
<td>564</td>
<td>8.66</td>
<td>195</td>
<td>55.58</td>
<td>398</td>
<td>8.66</td>
<td>1955</td>
<td>8.57</td>
<td>257</td>
<td>30.28</td>
</tr>
<tr>
<td></td>
<td>202</td>
<td>9.59</td>
<td>558</td>
<td>8.48</td>
<td>195</td>
<td>55.52</td>
<td>384</td>
<td>8.48</td>
<td>1862</td>
<td>8.02</td>
<td>257</td>
<td>27.98</td>
</tr>
<tr>
<td></td>
<td>195</td>
<td>9.55</td>
<td>552</td>
<td>8.05</td>
<td>175</td>
<td>9.99</td>
<td>366</td>
<td>8.05</td>
<td>1759</td>
<td>9.99</td>
<td>257</td>
<td>27.98</td>
</tr>
<tr>
<td></td>
<td>188</td>
<td>9.52</td>
<td>547</td>
<td>7.80</td>
<td>167</td>
<td>6.64</td>
<td>350</td>
<td>8.43</td>
<td>1098</td>
<td>5.4</td>
<td>257</td>
<td>24.68</td>
</tr>
</tbody>
</table>

TABLE IV
DISTRIBUTION OF THE DRUG-EVENT COUPLES AMONG DIFFERENT RELATIVE RISK INTERVALS AND AS A FUNCTION OF THE NUMBER OF COMBINATIONS TAKEN INTO ACCOUNT

<table>
<thead>
<tr>
<th>Measures</th>
<th>Number of the most highly ranked combinations taken into account</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBAM</td>
<td>10</td>
</tr>
<tr>
<td>EBp</td>
<td>(219)</td>
</tr>
<tr>
<td>SPRT$_2$</td>
<td>0.0</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>0.0</td>
</tr>
<tr>
<td>Poisson</td>
<td>(106)</td>
</tr>
<tr>
<td>$IC_{Gould}$</td>
<td>2.5</td>
</tr>
<tr>
<td>$IC_{Bate}$</td>
<td>0.0</td>
</tr>
<tr>
<td>$PRR$</td>
<td>2.4</td>
</tr>
<tr>
<td>ROR</td>
<td>2.2</td>
</tr>
</tbody>
</table>

TABLE V
AVERAGE OF THE FALSE POSITIVE PERCENTAGES FOR ALL THE DRUG-EVENT COMBINATIONS

<table>
<thead>
<tr>
<th>Measures</th>
<th>Number of the most highly ranked combinations taken into account</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBAM</td>
<td>10</td>
</tr>
<tr>
<td>EBp</td>
<td>(219)</td>
</tr>
<tr>
<td>SPRT$_2$</td>
<td>0.0</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>0.0</td>
</tr>
<tr>
<td>Poisson</td>
<td>(106)</td>
</tr>
<tr>
<td>$IC_{Gould}$</td>
<td>67.2</td>
</tr>
<tr>
<td>$IC_{Bate}$</td>
<td>0.0</td>
</tr>
<tr>
<td>$PRR$</td>
<td>64.8</td>
</tr>
<tr>
<td>ROR</td>
<td>64.9</td>
</tr>
</tbody>
</table>

of couples, thus underlining the better ranking performances for these measures, even with many couples.

Measure performances on drug–event couples with a low number of reports are presented in Table VII. Performances of EBAM, EBp, SPRT$_2$, $\chi^2$, Poisson, and $IC_{Bate}$ are more sensitive to an increase in the reports number when $a \leq 4$. The differences between these measures and the others are more significant for a low number of combinations (10 to 100) and especially when $a \leq 3$ or $a \leq 4$. In these cases, the measures derived from DuMouchel’s model (EBAM and EBp), $IC_{Bate}$, and SPRT$_2$ are the most efficient and provides comparable results.

Results according to the seriousness of events and background incidence are not presented here. In fact, the influence of these parameters on the results is not significant and the performances are comparable to those for the whole dataset (Table V).

IV. DISCUSSION

All the results presented in this paper were obtained with simulated data. Consequently, any conclusions have to be carefully transposed to the “real” domain of pharmacovigilance, especially with regard to performances of methods according to million) and lead to low proportions of false positive signals. Significant differences can be observed for a large number of couples (500), where the $\chi^2$ method becomes the least effective measure. This corroborates the relatively bad performances of $\chi^2$ according to ROC curves of Fig. 2. EBAM, EBp, Poisson, SPRT$_2$ and $IC_{Bate}$ are less sensitive to the increase in the number of couples, thus underlining the better ranking performances for these measures, even with many couples.

Measure performances on drug–event couples with a low number of reports are presented in Table VII. Performances of EBAM, EBp, SPRT$_2$, $\chi^2$, Poisson, and $IC_{Bate}$ are more sensitive to an increase in the reports number when $a \leq 4$. The differences between these measures and the others are more significant for a low number of combinations (10 to 100) and especially when $a \leq 3$ or $a \leq 4$. In these cases, the measures derived from DuMouchel’s model (EBAM and EBp), $IC_{Bate}$, and SPRT$_2$ are the most efficient and provides comparable results.

Results according to the seriousness of events and background incidence are not presented here. In fact, the influence of these parameters on the results is not significant and the performances are comparable to those for the whole dataset (Table V).

IV. DISCUSSION

All the results presented in this paper were obtained with simulated data. Consequently, any conclusions have to be carefully transposed to the “real” domain of pharmacovigilance, especially with regard to performances of methods according to
the SRS parameters. It seems especially important to study the influence of the choice of the model parameters’ values on the performances of the signal generation methods. This would evaluate the robustness of the results presented here. The qualitative knowledge exploited for the model definition is consensual knowledge found in the literature and expressed by pharmacovigilance experts. Moreover, the temporal evolution of the reporting probability (see (1) and [8]) follows the expected ones according to this knowledge. Therefore simulated data could be considered as realistic. However, an effort should be made in order to make the distributions of the marginal numbers of Table I comparable with the real ones for a comparable size subset of the real pharmacovigilance database. First, these numbers are exploited by the signal generation methods to determine the expected numbers of reports. Second, studying the distributions of these numbers is, to our knowledge, the only manner to quantitatively evaluate the SRS model at the database scale. Differences between simulated and real datasets can be observed, especially concerning the distribution of the number of drug–event combinations as a function of the number of reports per combination. In the dataset used by Van Puijenbroek et al. [3], the proportion of combinations with one, two, three and four or more reports was 68.4%, 14.2%, 6.2% and 11.2%, respectively, compared to 68.4%, 14.2%, 6.2% and 11.2%, respectively, compared to 68.4%, 14.2%, 6.2% and 11.2%, respectively, in our simulated data. On the other hand, DuMouchel [2] used a dataset that contained 35.5% of drug–event couples with one report. This demonstrates the disparity of “real” database features and, when performing stratification of the databases according to sex, age, etc., the disparity from a stratum to another. This makes the need to go further into the quantitative evaluation of the model of secondary importance. Eventually, some SRS features have not been taken into account in the model of the spontaneous reporting system. The most important one is the drugs interactions i.e., the fact that some events can be caused by the simultaneous exposure to two or more drugs and not by the drugs taken alone. DuMouchel proposes a method to identify the associations between an event and more than one drug by means of the “all-two-factor” model [15]. In order to evaluate such a method, it seems necessary to model drugs interactions in the simulated datasets.

By choosing the evaluation criteria of the present paper, we deliberately chose to evaluate measures of association and not automatic signal generation methods. However, our results also throw light on the generation criteria. Overall, there is one group including EBAM, EBp, SPRT, $\chi^2$, Poisson, and IC$_{Bate}$, and another including IC$_{Gould}$, PRR, ROR, and Yule’s $Q$. The former provides better results, especially for drug–event associations with a low number of reports. Such results are of major interest as drug–event couples with fewer than five reports are the most numerous in the database.

In the present study, IC is very sensitive to the choice of the prior parameter values. Bate et al. [1] and Orre et al. [14] propose using available data to define prior parameters. Gould [7] proposes fixed priors once for all the drug–event couples in the database. According to the results of the present study, exploiting prior knowledge derived from the data as Bate et al. do seems to provide better results. Such differences of results should be reduced when using larger databases; i.e., when the total number of reports in the database is more important (see Appendix). However, if surveillance strategies consist in applying signal generation methods to data subsets (serious events and/or recent drugs for example), the total number of reports of these subsets could be comparable to one of the simulated datasets used in the present study. Thus, exploiting available data in order to define priors seems a good solution. It is especially the solution adopted in DuMouchel’s method. In fact, EBAM is an empirical Bayesian method in the sense that priors parameters values are obtained by a maximum likelihood estimation using the data (see Appendix). The Bayesian approach seems to be efficient for pharmacovigilance database screening, as it provides a better estimation of the measures values for drug–event couples with a low number of reports.

Over the 1000 simulated datasets, 219(SD = 52) and 106(SD = 4) drug–event couples were placed equal first with EBp and Poisson, respectively. However, there was a very low number of false positive signals among the couples equally placed. For these two measures, the value corresponding to the number of reports of these subsets could be comparable to the one of the simulated datasets used in the present study. Thus, exploiting available data in order to define priors seems a good solution. It is especially the solution adopted in DuMouchel’s method. In fact, EBAM is an empirical Bayesian method in the sense that priors parameters values are obtained by a maximum likelihood estimation using the data (see Appendix). The Bayesian approach seems to be efficient for pharmacovigilance database screening, as it provides a better estimation of the measures values for drug–event couples with a low number of reports.

<table>
<thead>
<tr>
<th>Measures</th>
<th>Data features</th>
<th>$a = 5$</th>
<th>$e = 3M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBAM</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>EBp</td>
<td>(219) 0.3</td>
<td>1.2</td>
<td>27.6</td>
</tr>
<tr>
<td>SPRT$_2$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Poisson</td>
<td>(106) 0.0</td>
<td>0.0</td>
<td>26.8</td>
</tr>
<tr>
<td>IC$_{Gould}$</td>
<td>0.3</td>
<td>0.7</td>
<td>1.3</td>
</tr>
<tr>
<td>IC$_{Bate}$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
</tr>
<tr>
<td>PRR</td>
<td>0.2</td>
<td>0.4</td>
<td>0.9</td>
</tr>
<tr>
<td>ROR</td>
<td>0.2</td>
<td>0.3</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of the most highly ranked combinations taken into account</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 20 50 100 200 500 100 200 500 100 200 500 100 200 500 100 200 500</td>
</tr>
</tbody>
</table>

TABLE VI: AVERAGE OF PERCENTAGES OF FALSE POSITIVES FOR HIGHLY REPORTED COMBINATIONS
TABLE VII

AVERAGE OF PERCENTAGE OF FALSE POSITIVES FOR COMBINATIONS WITH A LOW NUMBER OF REPORTS

<table>
<thead>
<tr>
<th>Measures</th>
<th>Number of reports</th>
<th>( a \leq 2 )</th>
<th>( a \leq 3 )</th>
<th>( a \leq 4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBAM</td>
<td>43.6</td>
<td>47.8</td>
<td>53.3</td>
<td>60.6</td>
</tr>
<tr>
<td>EBp</td>
<td>43.6</td>
<td>47.8</td>
<td>53.3</td>
<td>60.4</td>
</tr>
<tr>
<td>SPRT(_2)</td>
<td>43.6</td>
<td>47.8</td>
<td>53.3</td>
<td>60.6</td>
</tr>
<tr>
<td>( \chi^2 )</td>
<td>51.5</td>
<td>54.0</td>
<td>59.8</td>
<td>67.2</td>
</tr>
<tr>
<td>Poisson</td>
<td>48.2</td>
<td>51.9</td>
<td>56.0</td>
<td>61.4</td>
</tr>
<tr>
<td>IC(_{\text{Gould}})</td>
<td>68.5</td>
<td>68.2</td>
<td>69.5</td>
<td>70.2</td>
</tr>
<tr>
<td>IC(_{\text{Bate}})</td>
<td>43.9</td>
<td>48.3</td>
<td>53.6</td>
<td>60.4</td>
</tr>
<tr>
<td>PRR</td>
<td>66.9</td>
<td>66.7</td>
<td>68.1</td>
<td>68.4</td>
</tr>
<tr>
<td>ROR</td>
<td>67.3</td>
<td>66.6</td>
<td>68.1</td>
<td>68.5</td>
</tr>
</tbody>
</table>

V. CONCLUSION

This paper demonstrates that it is possible to evaluate the automatic signal generation methods proposed in the literature within the field of pharmacovigilance. First, the realistic simulated datasets provided by modeling the spontaneous reporting system make it possible to estimate the performances of the methods. Second, the practical point of view adopted in this paper, which consists in ranking drug–event couples and in comparing the association measures by means of the percentage of false positive signals among a given number of the most highly ranked combinations, make it possible to identify the most efficient and applicable measures for users; i.e., the pharmacovigilance experts. According to this study, EBAM and \( \chi^2 \) provide the better results. These two methods use Bayesian measures exploiting available data for priors definition, thus proving the efficiency of this philosophy. However, their theoretical background and implementation are less obvious than other methods, like SPRT\(_2\) and \( \chi^2 \), which seem to provide good results as well.

The practical approach adopted in this paper for evaluating automatic signal generation methods implicitly involves the human, material, and time resources available for interpreting and exploiting retrieved information. It is of major interest, since the quantity of data now available is huge. Moreover, this approach should be adopted not only for evaluating measures of association, but also for developing new measures.

APPENDIX I

ASSOCIATION MEASURES AND SIGNAL GENERATION CRITERIA

1) Chi Square (Yates Correction): Chi square with Yates correction is defined as

\[
\chi^2 = \sum_{\text{4 cases of Table } 1} \frac{(a - E) - \frac{1}{2}}{E}.
\]  

The generation criterion is

\[
Pr\{\chi^2 > \chi^2\} \leq 0.05.
\]  

2) Proportional Reporting Ratio, PRR: PRR is defined as

\[
PRR = \frac{a \cdot d}{b \cdot c}.
\]

PRR was initially proposed for signal generation in pharmacovigilance by Evans et al. [9]. Criterion to generate a signal is \( PRR \geq 2 \) and \( \chi^2 \geq 4 \) and \( a \geq 3 \). Evans et al. calculate a confidence interval around the PRR as an alternative to chi-square. Van Puijenbroek et al. [3] generate a signal when the lower bound of the 95% confidence interval, noted \( LI_{95}(PRR) \), exceeds 1, with:

\[
LI_{95}(PRR) = \exp\{\ln(PRR) - 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\}.
\]  

3) Reporting Odds Ratio, ROR:

\[
ROR = \frac{a \cdot d}{b \cdot c}.
\]

Van Puijenbroek et al. propose a similar signal generation criterion as in PRR, i.e., \( LI_{95}(ROR) > 1 \) [3], with:

\[
LI_{95}(ROR) = \exp\{\ln(ROR) - 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\}.
\]
In a recent paper, Rothman et al. [16] argue that the ROR is a better measure than PRR for signal generation because it is a better estimator of the Risk Ratio (RR). They propose that its computation should be done by selecting control events; i.e., events independent of the exposure to the drug of interest. However, they mention the difficulty of introducing this extrinsic knowledge. They also notice that the lack of knowledge about the reporting ratio, which varies with many factors (seriousness and incidence of the events, exposure to the drug, etc.), makes the interpretation of the PRR and ROR equally problematic. As a response to the paper of Rothman, Waller et al. [6] propose that PRR and ROR are both disproportionality measures that give similar results when \(a\) and \(c\) (see Table I) are small compared with \((a + b)\) and \((c + d)\), respectively, as is the case in pharmacovigilance databases. Waller et al. also that in the context of pharmacovigilance, neither the PRR nor the ROR is meant to estimate the RR, but to rather identify suspicious drug–event couples. These measures should be evaluated according to their sensitivity, specificity, and predictive values obtained with a real dataset, and should not be compared with the true RR. Unfortunately, dataset that would permit such an evaluation does not yet exist as the true status of the drug–event relations in databases are unknown (this justifies the approach of the present paper).

4) Yule’s Q: Yule’s Q measure is defined as

\[
Q = \frac{ROR - 1}{ROR + 1} = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c}.
\]  

(9)

For Yule’s Q, the generation criterion is \(LI_{95} (Q) > 0\) [3], and:

\[
LI_{95} (Q) = Q - 1.96 \cdot \left( \frac{1}{2} \cdot (1 - Q^2) \cdot \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \right)
\]  

(10)

5) Poisson: This method assumes that for a drug–event couple, the number of reports is Poisson-distributed. A signal is generated if [3]

\[
Pr(\text{Pois}(E) > a) \leq 0.05
\]  

(11)

i.e.,

\[
1 - \sum_{k=1}^{a} \frac{e^{-E} \cdot E^k}{k!} \leq 0.05.
\]  

(12)

6) Sequential Probability Ratio Test, SPRT: SPRT was proposed by Spiegelhalter et al. [12] for monitoring the cumulative occurrence of various adverse clinical outcomes, and its use has been suggested by Evans in the context of pharmacovigilance [13]. By assuming a Poisson distribution for the number of reports and an event relative risk for “true” associations twice (SPRT₂) the event relative risk for coincidental associations, the SPRT₂ criterion for a signal generation is

\[
\ln(2) \cdot a - E \geq \ln(B)
\]  

(13)

where

\[
B = \ln \left( \frac{1 - \beta}{\alpha} \right)
\]  

(14)

\(\alpha\) and \(\beta\) being the risk of generating a false positive signal and a false negative signal, respectively. By choosing \(\alpha = \beta = 0.05\), we obtain the following generation criterion:

\[
\ln(2) \cdot a - E \geq 2.94.
\]  

(15)

7) Information Component, IC: Using a Bayesian approach, Bate et al. [1] define the IC as follows:

\[
IC = \log_2 \left( \frac{r}{p \cdot q} \right)
\]  

(16)

\(p, q, \) and \(r \) are, respectively, the probabilities of being exposed to the drug, of observing the considered event, and of having the drug–event association, given the data in Table I. \(p, q, \) and \(r \) are taken to be beta-distributed. Two sets of prior parameters have been proposed in literature.

a) Orre et al. [14] and Bate et al. [1] propose the non-informative prior distributions Beta\((p_1^0 = 1, p_2^0 = 1)\) and Beta\((q_1^0 = 1, q_2^0 = 1)\) for \(p\) and \(q\), respectively. Then, they consider the fact that the posterior expectation (Post-Expect) of the information component should tend toward zero when the observed number of reports tends toward zero. Consequently, they define the prior distribution for \(r\) as Beta\((r_1^2 = 1, r_2^2 = 1)\) Post-Expect\((p) \cdot Post-Expect(q) - 1\). So they introduce knowledge on the observed data in the prior parameters, and thus define a sort of semiempirical Bayesian method; b) Gould [7] proposes the same priors for \(p\) and \(q\), but do not exploit the available information for determining the prior parameters for \(r\). For \(r\), the prior distribution is Beta\((r_1^3 = 1, r_2^3 = 1)\); i.e., Beta\((r_1^3 = 1, r_2^3 = 3)\).

Therefore, the posterior distributions of \(p, q, \) and \(r \) are Beta distributions with parameters:

\[ p_1 = p_1^0 + (a + b) = 1 + a + b; \]

\[ p_2 = p_2^0 + N - (a + b) = 1 + N - a - b; \]

\[ q_1 = q_1^0 + (a + c) = 1 + a + c; \]

\[ q_2 = q_2^0 + N - (a + c) = 1 + N - a - c; \]

\[ r_1 = r_1^0 + a = 1 + a; \]

\[ r_2 = r_2^0 + N - a. \]

Then, by assuming a normal distribution for IC, Bate et al.’s signal generation criterion is \(LI_{95} (IC) > 0\), with:

\[
LI_{95} (IC) = \text{Expect}(IC) - 1.96 \cdot \text{SD}(IC)
\]  

(17)

where Expect and SD are the posterior expectation and standard deviation, respectively, with [7]

\[
E(IC) = \frac{1}{\ln(2)} \cdot (\Psi(r_1) - \Psi(r_1 + r_2) - (\Psi(p_1) - \Psi(p_1 + p_2) + \Psi(q_1) - \Psi(q_1 + q_2)))
\]  

(18)
(SD( IC))^2 = (ln(2))^{-2} \cdot (\Psi'(r_1) - \Psi'(r_1 + r_2) \\
+ (\Psi'(p_1) - \Psi'(p_1 + p_2) \\
+ \Psi'(q_1) - \Psi'(q_1 + q_2))).(19)

Ψ and Ψ’ being the digamma and the trigamma functions, respectively. These functions are tabulated in statistical software.

8) Empirical Bayes Method: DuMouchel [2] assumes a Poisson distribution with a mean μ for the number of reports related to a drug–event couple. Then he considers the rate of a drug–event couple. An a priori mixture of two gamma distributions is assumed for λ

λ_{apriori} \sim P \cdot \Gamma(a_1, b_1) + (1 - P) \cdot \Gamma(a_2, b_2). (20)

This mixture means that the probabilities that λ stems from a gamma distribution with parameters (a_1, b_1) and (a_2, b_2) are P and (1 - P), respectively. At the dataset scale, P and (1 - P) can be interpreted as the proportions of drug–event couples that stem from gamma distribution with parameters (a_1, b_1) and (a_2, b_2), respectively. The “empirical” character of the method comes from the estimation of the prior distribution parameters, q = \{ P, a_1, b_1, a_2, b_2 \}, by means of a maximum likelihood estimation from the data. The posterior distribution of λ is also a mixture of two gamma distributions

λ_{aposteriori} \sim Q \cdot \Gamma(a_1 + a, b_1 + E) \\
+ (1 - Q) \cdot \Gamma(a_2 + a, b_2 + E). (21)

Q is the posterior probability (see [2], [7] for more details on the computations). It is then possible to obtain the exact posterior mean of λ, termed EBAM. DuMouchel uses this value (in fact, DuMouchel uses the geometric mean derived from \log_2(λ)) to order the drug–event couples, and does not recommend a threshold. Gould [7] applies the same decision criterion as Bate et al. by computing the lower bound of the 95% confidence interval. This lower bound can be approached with a predefined precision. Indeed, Gamma quantiles are tabulated in commercial software, and a basic optimization procedure can easily find the quantiles of the posterior mixture. The quantile l corresponding to a given probability pr is the one that minimizes the following expression:

f = \left( Q \cdot \int_0^l \Gamma(x, a_1 + a, b_1 + E) \, dx \\
+ (1 - Q) \cdot \int_0^l \Gamma(x, a_2 + a, b_2 + E) \, dx \right) - pr. (22)

9) Alternative Generation Criterion for the Empirical Bayes Method: In the present paper, we propose an alternative signal generation criterion exploiting the empirical Bayes model of DuMouchel. As stated by the model in 8), the ratio λ, for each drug–event couple of the database, is supposed to stem from component 1 or 2 with the probability P and 1 - P, respectively.

For the definition of our new measure, we distinguish the two mixture components and consider the probability EBp that the λ ratio stems from a gamma distribution with an average exceeding 1. In fact, the higher this probability is, the higher the probability that λ itself exceeds 1 is. A drug–event couple associated with λ > 1 has an observed number of reports greater than the expected one and is supposed to be identified as a suspicious couple.

EBp is defined as

EBp = Q \cdot \delta_1 + (1 - Q) \cdot \delta_2 (23)

with $\delta_k = 1$ if the posterior mean of the component k exceeds 1, and $\delta_k = 0$ otherwise. With the notations previously stated

$k \in \{ 1, 2 \} \quad \begin{cases} \delta_k = 1, & \text{if } \frac{\mu + \sigma^2}{\delta_k + E} > 1 \\ \delta_k = 0, & \text{otherwise}. \end{cases} (24)$

A signal is generated when EBp exceeds a given value. In the present paper, we generate a signal when EBp > 0.5.

REFERENCES


Emmanuel Roux received the Ph.D. degree in automation and computer science for industrial and human systems from the University of Valenciennes, France, in 2002. He has worked for one year as a Postdoctoral Fellow in the Biostatistics Division of the INSERM U.472-IFR 69, Villejuif, France. His is currently a Postdoctoral Fellow in the Laboratory of Image and Signal Processing (LTSI), INSERM U.642, Rennes, France.

Frantz Thiessard received the Ph.D. degree in pharmacology from the University Victor Segalen of Bordeaux, France, in 2004. He specializes in public health and medical in forms, and is working on automatic signal generation on the French pharmacovigilance database.

Annie Fourrier received the Pharm.D. and Ph.D. degrees from the School of Public Health at Tulane University, New Orleans, LA. She is a Research Scientist in the Department of Pharmacology of the University Victor Segalen of Bordeaux in France. Her research expertise is in the evaluation of post-approval risk of medicines. She has conducted pharmacoepidemiological studies concerning the risk of vaccines and the association between drug consumption and cognitive decline in the elderly. She is currently involved in a project on signal detection in pharmacovigilance, and is a member of the French National Committee on Pharmacovigilance.

Bernard Begaud is a Professor of Pharmacology and Pharmacoepidemiology. His research expertise concerns statistical modeling in pharmacoepidemiology and the public health impact of drugs.

Pascale Tubert-Bitter is a Researcher at the French National Institute of Health and Medical Research in the Epidemiology and Biostatistics Department, Villejuif, France. She is interested in statistical methods for pharmacovigilance and clinical trials. She has published research papers in biometry, biostatistics, and epidemiology journals.