frontiers in Pharmacology

ORIGINAL RESEARCH
published: 18 September 2018
doi: 10.3389/fphar.2018.01010

Propensity Score-Based Approaches in High Dimension for Pharmacovigilance Signal Detection: an Empirical Comparison on the French Spontaneous Reporting Database

Émeline Courtois[1], Antoine Pariente[2], Francesco Salvo[2], Étienne Volatier[1], Pascale Tubert-Bitter[1†] and Ismaïl Ahmed[1†]*

[1] Biostatistics, Biomathematics, Pharmacoepidemiology and Infectious Diseases, INSERM, UVSQ (Université Paris-Saclay), Institut Pasteur, Villejuif, France, [2] Bordeaux Population Health Research Center, Pharmacoepidemiology Team (UMR 1219), INSERM, University of Bordeaux, Bordeaux, France

OPEN ACCESS

Edited by:
Marie-Christine Jaulent, Institut National de la Santé et de la Recherche Médicale (INSERM), France

Reviewed by:
Robert L. Lins, Retired, Antwerpen, Belgium
Juergen Landes, Ludwig-Maximilians-Universität München, Germany

***Correspondence:**
Émeline Courtois
emeline.courtois@inserm.fr

[†] These authors share last authorship

Specialty section:
This article was submitted to Pharmaceutical Medicine and Outcomes Research, a section of the journal Frontiers in Pharmacology

Received: 29 November 2017
Accepted: 20 August 2018
Published: 18 September 2018

Citation:
Courtois É, Pariente A, Salvo F, Volatier É, Tubert-Bitter P and Ahmed I (2018) Propensity Score-Based Approaches in High Dimension for Pharmacovigilance Signal Detection: an Empirical Comparison on the French Spontaneous Reporting Database. Front. Pharmacol. 9:1010. doi: 10.3389/fphar.2018.01010

Classical methods used for signal detection in pharmacovigilance rely on disproportionality analysis of counts aggregating spontaneous reports of a given adverse drug reaction. In recent years, alternative methods have been proposed to analyze individual spontaneous reports such as penalized multiple logistic regression approaches. These approaches address some well-known biases resulting from disproportionality methods. However, while penalization accounts for computational constraints due to high-dimensional data, it raises the issue of determining the regularization parameter and eventually that of an error-controlling decision rule. We present a new automated signal detection strategy for pharmacovigilance systems, based on propensity scores (PS) in high dimension. PSs are increasingly used to assess a given association with high-dimensional observational healthcare databases in accounting for confusion bias. Our main aim was to develop a method having the same advantages as multiple regression approaches in dealing with bias, while relying on the statistical multiple comparison framework as regards decision thresholds, by considering false discovery rate (FDR)-based decision rules. We investigate four PS estimation methods in high dimension: a gradient tree boosting (GTB) algorithm from machine-learning and three variable selection algorithms. For each (drug, adverse event) pair, the PS is then applied as adjustment covariate or by using two kinds of weighting: inverse proportional treatment weighting and matching weights. The different versions of the new approach were compared to a univariate approach, which is a disproportionality method, and to two penalized multiple logistic regression approaches, directly applied on spontaneous reporting data. Performance was assessed through an empirical comparative study conducted on a reference signal set in the French national pharmacovigilance database (2000–2016) that was recently proposed for drug-induced liver injury. Multiple regression approaches performed better in detecting true positives and false positives. Nonetheless, the performances of the PS-based methods using

matching weights was very similar to that of multiple regression and better than with the univariate approach. In addition to being able to control FDR statistical errors, the proposed PS-based strategy is an interesting alternative to multiple regression approaches.

Keywords: pharmacovigilance, signal detection, propensity score in high dimension, spontaneous reports, penalized multiple regression, FDR

1. INTRODUCTION

Once a drug is introduced on the market, many people are exposed to it in real-life conditions, which can be very different from those evaluated in clinical trials. The goal of pharmacovigilance is to detect as quickly as possible potential adverse reactions which could be induced by drug exposure. To achieve this challenging task, health authorities collect and monitor spontaneous reports of suspected adverse events (AEs), mainly from practitioners. In France, such a pharmacovigilance database is maintained by the National Agency for the Safety of Drugs and Health Products (*Agence Nationale de Sécurité du Médicament et des Produits de Santé*, ANSM). It contained around 431,000 reports at the end of July 2016. In recent years, about 36,000 reports have been reported annually.

To exploit this amount of data, several automated signal detection tools have been developed. The term signal refers to a (drug, AE) pair whose association is highlighted by a signal detection method. The aim of such methods is to draw attention to unexpected associations by acting as hypothesis generators. The signals thus generated must be further investigated by pharmacovigilance experts in order to draw definite conclusions. The most common methods used by health agencies are disproportionality methods (Almenoff et al., 2007). These methods rely on a three-step strategy: for each (drug, AE) pair (1) determine the number of reports involving this specific pair; (2) construct a measure of the degree to which the observed count of reports exceeds the expected count if independence applies (the disproportionality measure); and (3) decide that a signal has occurred if this measure exceeds a specified threshold value. The procedure is conducted for all (drug, AE) pairs simultaneously (Ahmed et al., 2015). More recently, these disproportionality methods were extended by integrating a false discovery rate (FDR) estimation procedure in order to take into account comparison multiplicity (Benjamini and Hochberg, 1995; Ahmed et al., 2010).

To address the shortcomings of disproportionality methods in dealing with confounding like masking effect and co-prescription bias (Harpaz et al., 2012), multiple logistic regression approaches have been proposed. While being more computationally intensive, they have been shown through empirical studies to be promising signal detection methods (Caster et al., 2010; Harpaz et al., 2013; Marbac et al., 2016; Ahmed et al., 2018). Instead of considering aggregated data as disproportionality methods do, these newer approaches analyse spontaneous reports directly: the observation becomes the individual report, the outcome is the presence/absence of a given AE, and the covariates are all drug presence indicators (Caster et al., 2010). Because of the very large number of potential covariates, a lasso version of the logistic regression has been used (Tibshirani, 1996). It consists in maximizing the log-likelihood of a classical multiple logistic regression model, penalized by a function proportional to the L_1 norm of the regression coefficients. This penalty makes it possible to shrink some coefficients associated with the covariates to exactly zero. Cross validation is commonly used to choose the penalty value and it achieves good performances in terms of prediction. However, determining the best penalization parameter in the variable selection framework is a challenging task.

Another way to deal with confounding bias issues is to summarize all the information into a propensity score (PS). PSs are classically used in epidemiological studies investigating one targeted association (i.e., one drug, one AE) to deal with confounding like indication bias. In recent years, this methodology has been used with high-dimensional health-care databases, like claims data, by collapsing all the numerous covariates present in these databases into a single value which is easier to manipulate (Seeger et al., 2005; Schneeweiss et al., 2009). Tatonetti et al. (2012) investigated the use of one PS strategy in the context of spontaneous reporting data and illustrated its interest by comparing it to a disproportionality method.

In this paper, we develop several PS-based approaches for signal detection from spontaneous reporting and compare them to penalized multiple regression approaches. We built the PSs with four different algorithms adapted to high dimensional settings and for each of them, we considered three different integration strategies to perform signal detection. Our proposed PS-based methods include a signal detection rule relying on FDR estimation. This comparison is performed with the French national pharmacovigilance database (2000–2016) using a large and recently published reference set pertaining to a common adverse reaction: drug-induced liver injury (DILI) (Chen et al., 2011, 2016).

2. METHODS

2.1. Reference Methods

The most common proposed methods for signal detection are disproportionality methods, which rely on the aggregated form of the data. These methods aim to detect higher-than-expected combinations of drugs and events in the database. Nevertheless, two kinds of biases are created by the univariate feature of disproportionality methods: masking and the co-reporting confounding effect (Harpaz et al., 2012). In order to address these two bias issues, other approaches in pharmacovigilance have

lately emerged that rely on multiple logistic regressions (Caster et al., 2010). Individual spontaneous reports are directly analyzed instead of aggregated counts of drug-event combinations. We denote by I the total number of drugs and J the total number of AEs in the database. Let Y_j indicate the presence or absence of AE_j, and X_i denote the presence or absence of $drug_i$. The corresponding logistic model for each AE_j is

$$\text{logit}(\Pr(Y_j = 1)) = \beta_0^j + \sum_{i=1}^{I} \beta_i^j X_i. \quad (1)$$

In spontaneous reporting databases, the number of drug covariates I is very large. To deal with this high-dimensional issue and in order to derive the most parsimonious model, a penalized lasso logistic regression was implemented (Tibshirani, 1996). This consists in maximizing the log-likelihood of model (1) minus a penalization term

$$pen_j(\lambda) = \lambda\, |\beta^j|_1 = \lambda \sum_{i=1}^{I} |\beta_i^j|. \quad (2)$$

Depending on the penalization parameter λ, the lasso regression shrinks most of the coefficients associated to the covariates to exactly zero, so these covariates are not retained in the model. Usually cross validation is used to select the best value of λ in terms of prediction error, but this does not achieve good performance in a variable selection framework which is that of signal detection. Here, we considered two alternative strategies to this penalization parameter selection: the BIC-Lasso and the class-imbalance subsampling lasso (CISL) (Ahmed et al., 2018).

2.1.1. BIC-Lasso

This method uses the Bayesian Information Criterion. First, a lasso regression is computed for a predefined set of penalization parameter values λ_ks. To each tested λ_k is associated a set of variables selected by the lasso regression. The BIC is then calculated from a classical logistic regression model for each of these sets:

$$\text{BIC} = -2\ln(L) + k\ln(N) \quad (3)$$

where L is the likelihood of the regression model, N the number of observations, and k the number of parameters in the model. We declared as signals all drugs positively associated with the outcome in the model that minimizes the BIC.

2.1.2. CISL

Another way to get around the penalization parameter selection issue is the stability selection method (Meinshausen and Bühlmann, 2010). To take into account the sparsity of pharmacovigilance databases with low frequencies of the outcomes, Ahmed et al. (2018) proposed a variation of this method: the CISL algorithm. In this method, samples are drawn following an nonequiprobable sampling scheme with replacement to allow a better representation of individuals who experienced the outcome of interest. For a given set of penalization parameter values, lasso logistic regressions are computed in each of these samples. The CISL procedure consists in computing the quantity:

$$\widehat{\pi}_i^b = \frac{1}{E} \sum_{\eta=1}^{E} \mathbf{1}[\widehat{\beta}_i^{\eta,b} > 0], \quad (4)$$

where E is the maximum value of covariates selected by all the lasso regressions, $\eta \in \{1, .., E\}$ is the number of predictors selected and $\widehat{\beta}_i^{\eta,b}$ is the regression coefficient estimated by the logistic lasso for drug i, on sample $b \in \{1, .., B\}$, for a model including η covariates. An empirical distribution of $\widehat{\pi}_i^b$ for each drug is obtained over all B samples. A covariate is finally selected by the CISL method if a given quantile of the distribution of $\widehat{\pi}_i^b$ is non-zero. In this work, we considered the covariate sets established with the 5% and the 10% quantiles of these distributions.

In the following we refer to these two methods as the multiple regression approaches.

2.2. Propensity Score Approaches

The PS is defined as the probability of being exposed to a treatment given the observed covariates (Rosenbaum and Rubin, 1983). Conditionally on the PS, treatment exposure and the observed covariates are independent: patients with similar PSs will have on average similar covariate distributions between the exposed and unexposed subjects. This relies on the strong assumption that there are no unmeasured confounders. This means that all variables that affect both treatment assignment and outcome, thus potentially inducing spurious associations, have been measured. The PS is used in the context of observational studies as a balancing score that allows for non-randomized trials to reproduce the conditions of a randomized experiment by making observed baseline characteristics comparable in two different treatment groups. Thus, it addresses confounding biases like indication bias. In this framework, the PS is unknown and has to be estimated from observed data, usually using a logistic regression. Each patient is then assigned a predicted probability of being exposed that is calculated from the PS regression.

2.2.1. Propensity Score Modeling

It is recommended to include predictors and confounders in the PS model i.e., covariates that are related to the outcome or to the outcome and the exposure. It is also strongly advised to avoid instrumental variables i.e., variables that are only related to the exposure (Patorno et al., 2013). In practice, it is sometimes recommended to make *a priori* variable selection according to expert knowledge (Brookhart et al., 2006). In the high-dimensional framework such as health-care databases where thousands of covariates are entered, this approach cannot hold. The high-dimensional propensity score (hdPS) algorithm was proposed to deal with this kind of difficulty (Schneeweiss et al., 2009). For a given exposure, the choice of candidate covariates to include in the PS model relies on empirical assessment of covariates prevalence and strength of association with the exposure and with the outcome of interest. The algorithm ranks all input variables by their potential for confounding and the

investigator must choose the top-ranked variables to be included in the model for estimating the PS.

In the present work we aimed at computing the PS for each drug in the database by selecting among all the other drugs those to be included in the PS estimation model. We implemented four PS-estimation methods. We considered three variable selection algorithms: hdPS and the two other variable selection methods described above: the BIC-Lasso and the CISL [with the positive constraint for β replaced by a non-zero constraint in Equation (4)]. Using the BIC-lasso methodology, each covariate associated with the smallest BIC models was selected. Using CISL, all covariates with a 10% percentile of its distribution greater than zero were selected. Using hdPS, the 20 top-ranked variables were selected. For the sake of completeness, we also implemented a PS-estimation method which relies on a machine-learning algorithm, as was recently proposed by Lee et al. (2010): we considered a gradient tree boosting (GTB) algorithm (Hastie et al., 2009).

2.2.2. Use of the Propensity Score

Once the PS is estimated, four methodologies may be used to remove confounding in estimating the treatment effect: (1) adjustment on the PS; (2) stratification on the PS; (3) matching on the PS; (4) weighting on the PS. Adjustment on the PS consists in considering the estimated PS as a covariate and including it as an additional variable, with the exposure, in the regression model on the outcome. Stratification on the PS is based on a population stratification according to the quantiles of the PS distribution. The treatment effect is then estimated in each subgroup of patients. Matching on the PS consists in matching one treated subject to one (or more) untreated subject which have similar values of their estimated PS. Classically, subjects are matched with a nearest neighbor matching algorithm within a specified caliper distance equal to 0.2 of the standard deviation of the estimated PS. Weighting on the PS amounts to assigning to each subject a weight that is calculated from the estimated PS. The treatment effect is estimated on the pseudo population which is built according to these weights (Austin, 2011).

In this study, we considered two of these four different PS methodologies. First we implemented an adjustment on the PS. For every (AE_j, $drug_i$) pair of interest, we computed the following logistic regression model:

$$\text{logit}(\Pr(Y_j = 1)) = \beta_0^j + \beta_i^j X_i + \widetilde{\beta}_i^j \widehat{e}_i \tag{5}$$

where $\widehat{e}_i$ is the estimated PS of $drug_i$ for all the observations.

We also implemented the inverse probability of treatment weighting (IPTW, Austin, 2011). For a given $drug_i$, the weights to be attributed to each observation are defined by

$$w_i^{IPTW} = \frac{X_i}{\widehat{e}_i} + \frac{1 - X_i}{1 - \widehat{e}_i}. \tag{6}$$

Because some drugs have been reported very rarely, and do not have many reports in common with AEs, matching subjects on PS may cause dramatic losses in terms of subjects who experienced both the exposure and the outcome. Keeping this constraint in mind, we performed another kind of weighting to mimic the classical matching by targeting the same estimand as pair-matching on the PS. The proposed weights are referred to the matching weights (MW) (Li and Greene, 2013; Franklin et al., 2017).

$$w_i^{MW} = \frac{\min(\widehat{e}_i, 1 - \widehat{e}_i)}{X_i \widehat{e}_i + (1 - X_i)(1 - \widehat{e}_i)}. \tag{7}$$

For these two weighting approaches, we then computed univariate weighted logistic regressions by considering w_i^{IPTW} or w_i^{MW}.

For all these PS-based approaches, the signal detection rule is the one applied in the hypothesis testing framework. To take the multiplicity of the tested hypotheses into account, we applied an FDR controlling procedure adapted to the one-sided null hypothesis setting that stands in pharmacovigilance (Ahmed et al., 2010). An FDR adjusted p-value is estimated for each comparison tested. A signal is considered to occur if this p-value corrected for multiplicity is below an FDR threshold.

3. MATERIAL AND COMPARISON SETTINGS

To assess the performances of our PS-based approaches, we compared their ability to correctly classify true and false signals to multiple regression approaches presented above and a univariate approach, using the French pharmacovigilance database on a recently proposed reference set pertaining to DILI adverse events.

3.1. Comparison Set-Up

We compared the ability of each method to detect true signals and not to detect false signals. For the BIC-Lasso approach, we declare as signals all drugs positively associated with the outcome in the model that minimizes the BIC. All drugs selected by CISL are considered as a signal since the positive association with the outcome is taken into account in the algorithm. For all PS-based approaches, a 5% level FDR controlling procedure was used to determine which drug-AE associations are considered as signals. We also included a method based on a simple univariate logistic regression: for each AE_j, $j \in \{1, .., J\}$ and each $drug_i$, $i \in \{1, .., I\}$ we computed:

$$\text{logit}(\Pr(Y_j = 1)) = \beta_0^j + \beta_i^j X_i. \tag{8}$$

This method was implemented to serve as a reference level. It can be assimilated to a reporting odds ratio method, which is a disproportionality method for signal detection (van Puijenbroek et al., 2002). We applied the same FDR controlling procedure to this univariate approach.

All the analyses were performed with R version 3.4.0 (R Core Team, 2017). All the logistic regressions were computed with the speedglm R package v0.3–2. All lasso regressions were implemented using the glmnet R package v2.0–10. GTB algorithms for PS estimation were implemented with the xgb.train function from the xgboost R package 0.6–4. The default values of the xgb.train function were used, except for the learning rate, which was fixed at 0.1.

Table 1 summarizes the main advantages and disadvantages of the different signal detection methods presented in this section. For the sake of clarity, we refer to the multiple regression approaches in the following as BIC-Lasso, CISL-5%, CISL-10% and to the univariate approach as Univ. The terms adjustPS-BIC, mwPS-BIC, iptwPS-BIC, adjustPS-CISL, mwPS-CISL, iptwPS-CISL, adjustPS-GTB, mwPS-GTB, iptwPS-GTB, and adjustPS-hdPS, mwPS-hdPS, iptwPS-hdPS refer to the PS approaches implemented, the way PS was estimated (BIC, CISL, GTB or hdPS) and how PS was taken into account (adjustment, MW, or IPTW).

3.2. The French Pharmacovigilance Database

The performances of our approaches were evaluated and compared by an empirical analysis using data from the French pharmacovigilance database. Drugs are coded according to the 5th level of the Anatomical Therapeutic Chemical (ATC) hierarchy, and the AEs to the Preferred Term (PT) level of the Medical Dictionary for Regulatory Activities (MedDRA).

We applied some filters to the data and considered (i) only products which are known to be a drug (leaving out vaccines, phytotherapy, homeotherapy, dietary supplement, oligotherapy and enzyme inhibitor); (ii) reactions which are targeted as AEs (overdoses or medication errors are not taken into account). Drugs are listed according to their active substances, which are encoded with the ATC classification and additional codes for substances that do not have a corresponding code. We extracted data over the period 2000–2016, which represents 382,484 reports with 5,906 different AEs and 2,344 different drugs.

3.3. Set of Reference Signals

The set of reference signals that we used to compare the performances of the methods is the DILIrank set recently established by Chen et al. (2011) which considers a specific adverse event: DILI. After determining a list of keywords related to the DILI event, this set was developed by text-mining the FDA-approved drug labels. A severity level was associated to each keyword. Drugs were classified into three DILI-related categories according to where keywords appear in the labeling section (Warnings & Precautions section, for example) of the FDA-approved drug labels, and the severity level of the involved keywords: most DILI concern, less DILI concern and no DILI concern. In 2016, this classification was refined by including information from other data sources to assess the causal relationship between each considered drug and a DILI event. Only drugs confirmed as DILI cause were retained (Chen et al., 2016).

We translated DILI labeling keywords into PT codes from the MedDRA classification: 91 PT codes were considered as DILI related. Each spontaneous report involving one of these 91 PT was thus considered as a reported DILI event, which translated into 22,440 DILI reports in the French pharmacovigilance database. To avoid numerical issues due to very low drug-reporting frequency, we limited the comparison to drugs which have more than three reports in common with a DILI, and have more than ten reports in total. This restriction applied, we considered 922 different drugs out of the 2,344 original drugs from the spontaneous reporting database. Considering these drugs, the final DILIrank set accounted for 417 (drug, DILI) pairs: 90 no DILI concern pairs, 213 less DILI concern pairs and 114 most DILI concern pairs. We considered as negative controls the no DILI concern pairs, and as positive controls the most DILI concern ones, which involve drugs associated with severe DILI outcomes.

4. RESULTS

Performances of the methods were assessed in terms of number of signals detected, number of true positives and number of false negatives identified, sensitivity and specificity, positive predictive value (PPV) and false discovery proportion (FDP). **Table 2** summarizes the results. AdjustPS-BIC, adjustPS-CISL, adjustPS-GTB, and adjustPS-hdPS detected more signals than the other PS-based approaches with 308, 275, 273, and 310 signals respectively. All the iptwPS-based approaches led to the lowest number of signals with 35, 63, 70, 34 signals for iptwPS-BIC, iptwPS-CISL, iptwPS-GTB, and iptwPS-hdPS. The mwPS-based approaches gave an intermediate number of detected signals with 147, 121, 136, and 139 signals for mwPS-BIC, mwPS-CISL, mwPS-GTB, and mwPS-hdPS respectively. Multiple regression approaches detected between 99 and 173 signals. The univariate approach detected the largest number of signals: 359.

Multiple regression approaches gave the highest positive predictive values (PPV): 96.97, 100, and 100% for BIC-Lasso,

TABLE 1 | Main advantages and disadvantages of the compared signal detection methods.

Methods	Advantages	Disadvantages
Univariate approach (disproportionality method)	Very fast computation time Detection threshold based on classical test theory (*p*-values, FDR correction)	Do not account for multiple exposures
Penalized multiple logistic regression methods	Fast computation time Account for multiple exposures	Detection threshold not relying on classical test theory
Propensity-score based methods	Account for multiple exposures Detection threshold based on classical test theory (*p*-values, FDR correction)	Long calculation time for the propensity score estimation step

TABLE 2 | Number of signals detected for each method.

Method	Number of generated signals	Number of signals with known status (positive or negative)	True positives signals			False positives signals		
			n	PPV	Sensitivity	*n*	FDP	Specificity
Univ	359	105	86	81.90	75.44	19	18.10	78.89
BIC-lasso	173	66	64	96.97	56.14	2	3.03	97.78
CISL-5%	99	43	43	100.00	37.72	0	0.00	100.00
CISL-10%	109	48	48	100.00	42.11	0	0.00	100.00
adjustPS-BIC	308	96	81	84.38	71.05	15	15.62	83.33
mwPS-BIC	147	53	52	98.11	45.61	1	1.89	98.89
iptwPS-BIC	35	14	13	92.86	11.40	1	7.14	98.89
adjustPS-CISL	275	86	75	87.21	65.79	11	12.79	87.78
mwPS-CISL	121	50	49	98.00	42.98	1	2.00	98.89
iptwPS-CISL	63	17	14	82.35	12.28	3	17.65	96.67
adjustPS-GTB	273	85	74	87.06	64.91	11	12.94	87.78
mwPS-GTB	136	52	49	94.23	42.98	3	5.77	96.67
iptwPS-GTB	70	28	25	89.29	21.93	3	10.71	96.67
adjustPS-hdPS	310	93	83	89.25	72.81	10	10.75	88.89
mwPS-hdPS	139	54	53	98.15	46.49	1	1.85	98.89
iptwPS-hdPS	34	16	15	93.75	13.16	1	6.25	98.89

For the BIC-lasso a signal is a positive association in the selected model, for CISL-5% and CISL-10% a signal is a selected variable obtained according to a distribution quantile (5% or 10%) in the CISL methodology. For Univ and all the PS-based approaches, a signal is an association FDR significant. The number of true positives signals is the number of positives controls detected by the method. The number of false positives signals is the number of negatives controls detected by the method.
PPV: Positive Predictive Value.
FDP: False Discovery Proportion.

CISL-5% and CISL-10% respectively. All the mwPS-based approaches gave similar results with a PPV around 98%, except for mwPS-GTB with a PPV equal to 94.23%. The univariate approach had the lowest PPV: 81.90%. AdjustPS-based approaches had slightly better performances: 84.38, 87.21, 87.06, and 89.25% for adjustPS-BIC, adjustPS-CISL, adjustPS-GTB, and adjustPS-hdPS respectively. Depending on the method used to estimate the PS, iptwPS-based approaches performed differently with PPVs ranging from 82.35% for iptwPS-CISL to 93.75% for iptwPS-hdPS.

Univ, adjustPS-BIC, adjustPS-CISL, adjustPS-GTB, and adjustPS-hdPS had the best performance in terms of sensitivity: between 64.91% for adjustPS-GTB and 75.44% for Univ, but they had the poorest specificity: between 78.9% for Univ and 88.9% for adjsutPS-hdPS. On the other hand, multiple regression approaches and mwPS-based approaches had very good specificity. BIC-Lasso, CISL-5% and CISL-10% had a specificity equal to 96.97, 100, and 100% respectively. MwPS-BIC, mwPS-CISL, mwPS-GTB, and mwPS-hdPS had a specificity equal to 98.11, 98.00, 94.23, and 98.15% respectively. Multiple regression approaches had sensitivity between 37.72% for CISL-5% and 56.14% for BIC-Lasso. MwPS-based approaches had sensitivity between 42.98% for mwPS-CISL and mwPS-GTB, and 46.49% for mwPS-hdPS. IptwPS-based approaches had good specificity but very low sensitivity.

Figure 1 shows the number of true positives or true negatives according to the number of signals generated by Univ, BIC-Lasso and mwPS-BIC, where signals are ranked according to their *p*-values for BIC-Lasso, and according to their adjusted *p*-values for Univ and mwPS-BIC. BIC-Lasso and mwPS-BIC were chosen to be represented here since they showed either better or similar performances to the other multiple regression or PS-based approaches respectively (**Figures S1**–**S3**). Regarding true positive signal detection, BIC-Lasso performed the best and Univ the worst, mwPS-BIC having a comparable performance to BIC-Lasso. Concerning false positive signal detection, mwPS-BIC performed the best since it detected the first false positive the latest. Univ detected far more signals than mwPS-BIC and BIC-Lasso, but its true positive detection rate was worse on its latest detected signals. Indeed, 36% of its first 150 generated signals were true positive controls compared to 15% of its last 209 signals. The same trend was observed for false positive detection: Univ detected 2% of false positives within its first 150 generated signals and about 7% over its last 209 signals.

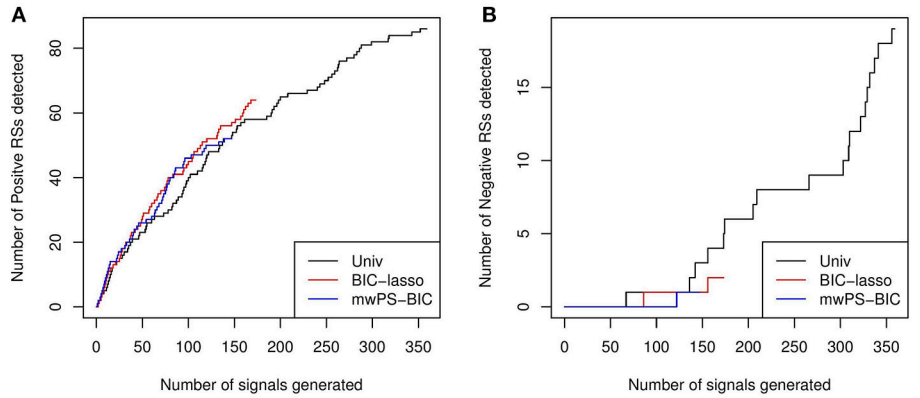

FIGURE 1 | (A) Number of positive reference signals detected according to number of signals generated by BIC-Lasso, mwPS-BIC and Univ, where signals are ranked in ascending order by their associated *p*-values for BIC-Lasso and by their adjusted *p*-values for mwPS-BIC and Univ. **(B)** Number of negative reference signals detected according to number of signals generated by BIC-lasso, mwPS-BIC and Univ, where signals are ranked in ascending order by their associated *p*-values for BIC-Lasso and by their adjusted *p*-values for mwPS-BIC and Univ.

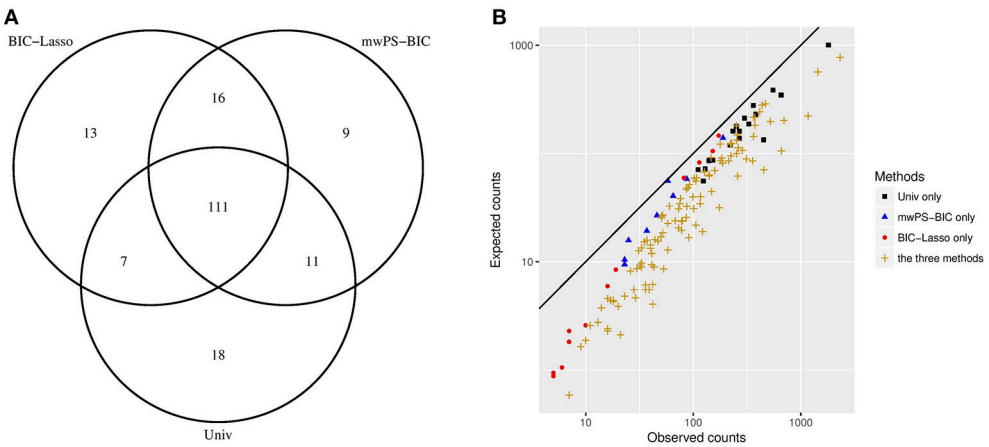

FIGURE 2 | (A) Distribution of the first 147 signals generated between Univ, BIC-Lasso and mwPS-BIC. **(B)** Observed counts vs. expected counts of signals generated by Univ only, BIC-Lasso only, mwPS-BIC only and by the three methods, considering their first 147 generated signals. Observed counts *n* are number of reports which involved the signal considered, expected counts *e* are those expected if independence applies between the drug and the AE that form the signal considered . They are calculated as follows: for a signal (drug$_i$, AE$_j$) $e_{i,j} = \frac{N_i \times N_j}{N}$ where N_i, N_j are the observed counts of drug$_i$ and AE$_j$ respectively, and N the total number of observations.

Looking more specifically at the PS-based approaches, they behaved similarly within a given PS-estimation method (**Figures S1**, **S2**). Considering the three PS methodologies, mwPS-based approaches performed better for an equal number of signals generated (**Figure S3**).

Figure 2A shows the overlap between the first 147 signals generated by Univ, BIC-Lasso and mwPS-BIC and **Figure 2B** the observed counts n vs. the expected counts e of some of these signals according to the method(s) by which they were detected. 111 signals were detected by all methods: **Figure 2B** shows that these signals are characterized by a greater observed risk ratio ($\frac{n}{e}$) and/or a greater support of the data (large n) in comparison to signals detected solely by one method. **Figure 2B** also shows that the methods tend to focus on different types of signals: in particular the BIC-Lasso highlighted some signals with relatively weak support of the data ($n < 10$) whereas Univ highlighted signals with low observed risk but a large number of reports. mwPS-BIC had an intermediary behavior.

5. DISCUSSION

Development of novel signal detection methods is crucial to improve the responsiveness and the efficiency of post-marketing surveillance systems. The main issue is to develop techniques that give a reasonable number of signals for further analysis by experts, while providing a list of relevant signals with the fewest possible false associations. Using spontaneous reporting

databases is difficult because of their high dimension and extreme imbalance. This study compared recently proposed multiple approaches and a set of methods based on the PS in a signal detection framework. To do so, we used a recently developed set of reference signals pertaining to liver injury: the DILIrank set.

The best performing PS-based approach is the one which relies on the MW. Its performance in terms of true/false signal detection, specificity and sensitivity is very close to that of the multiple regression approaches, which performs the best. It also generates roughly the same number of signals as BIC-Lasso and CISL, the latter being conservative for highly reported adverse events, as is the case for DILI (Ahmed et al., 2018). The mwPS-based approaches have an intermediary behavior between multiple regression approaches and univariate approach in the type of signal generated.

In accordance with the literature, adjustment on the PS does not achieve the best performances (Stuart, 2010). The adjustment approaches had similar behavior to the univariate approach. In particular, they generated very large numbers of signals in comparison to the other PS-based approaches and multiple regression approaches, and they had poor specificity and good sensitivity.

The iptwPS-based approaches performed extremely poorly. Unlike the MW, weights from IPTW are not normalized and could be very large for untreated individuals with low PS. This numerical instability due to high weights with IPTW weighting has already been reported (Yoshida et al., 2017). To avoid this issue, a solution is to do weight-trimming: weights greater than a given value are assigned this value (Seeger et al., 2017).

In addition to the well-known variable selection algorithm in PS estimation, hdPS, here we implemented three other PS-estimation methods: two based on variable selection algorithms, BIC-Lasso or CISL, and one derived from a machine-learning algorithm: GTB. A drawback with the hdPS algorithm in our setting is that the PS obtained with this method is AE dependent: for each drug to be included in the PS, its association with the AE under study has to be computed. This can become very time-consuming when screening several thousands of AEs. On the contrary, the three other PS-estimation methods have a computational advantage: once the PS is estimated for a given drug, it can be used to test associations with any AEs. When PS is estimated with the covariates selected by BIC-Lasso, CISL or with a GTB algorithm, PS-based approaches are competitive with multiple regression approaches in terms of calculation time. Regressions like weighted univariate logistic regressions can be easily performed. Depending on the PS methodology used (adjustment or weighting), results are roughly comparable according to the way PS is estimated.

A major constraint in developing signal detection methods is to obtain a reliable set of reference signals to assess performances. Here we used a recently established set, the DILIrank set pertaining to a common adverse event, which is not the case for all the AEs in the database. Furthermore, three levels of DILI risk assessment are defined in DILIrank for drugs: most DILI concern, less DILI concern and no DILI concern. We chose to consider only most DILI concern drugs as true positive controls, leaving less DILI concern drugs, because we decided to focus on drugs which could lead to severe DILI related outcomes.

The results suggest that the proposed PS-based methodology is an interesting complement to other existing methods. In a sense, it combines the main strengths of both univariate and multiple regression approaches: it makes it possible to account for co-reported drugs while using multiple hypothesis testing theory as regards the detection threshold. Further studies on alternative reference sets and simulation studies will be useful to confirm its potential for automated signal detection in pharmacovigilance.

AUTHOR CONTRIBUTIONS

ÉC, IA, and PT-B conceived and designed the study. ÉV contributed to the design of the work. AP and FS contributed to the acquisition and the interpretation of data for the work. ÉC performed the computations. ÉC, IA, PT-B, and ÉV discussed the results. ÉC drafted the manuscript with support from IA and PT-B. All authors critically revised the work and approved the final manuscript.

ACKNOWLEDGMENTS

The authors thank the regional pharmacovigilance centers and the ANSM for providing the pharmacovigilance database dataset.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2018.01010/full#supplementary-material

REFERENCES

Ahmed, I., Dalmasso, C., Haramburu, F., Thiessard, F., Broët, P., and Tubert-Bitter, P. (2010). False discovery rate estimation for frequentist pharmacovigilance signal detection methods. *Biometrics* 66, 301–309. doi: 10.1111/j.1541-0420.2009.01262.x

Ahmed, I., Pariente, A., and Tubert-bitter, P. (2018). Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions. *Statist Methods Med. Res.* 27, 785–797. doi: 10.1177/0962280216643116

Ahmed, I., Bégaud, B., and Tubert-Bitter, P. (2015). "chapter 13: evaluation of post-marketing safety using spontaneous reporting databases," in *Statistical Methods for Evaluating Safety in Medecine Product Development*, ed A. L. Gould (Chichester, UK: Wiley), 332–344.

Almenoff, J.S., Pattishall, E.N., Gibbs, T.G., Dumouchel, W., Evans, S. J., and Yuen, N. (2007). Novel statistical tools for monitoring the safety of marketed drugs. *Clin. Pharmacol. Therapeut.* 82, 157–166. doi: 10.1038/sj.clpt.6100258

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational Studies. *Multivar. Behav. Res* 46. 399–424. doi: 10.1080/00273171.2011.568786

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* 57, 289–300.

Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *Am. J. Epidemiol.* 163, 1149–1156. doi: 10.1093/aje/kwj149

Caster, O., Norén, G. N., Madigan, D., and Bate, A. (2010). Large-Scale regression-based pattern discovery: the example of screening the WHO global drug safety database. *Statist. Anal. Data Minig* 3, 197–208. doi: 10.1002/sam.10078

Chen, M., Suzuki, A., Thakkar, S., Yu, K., and Hu, C. (2016). DILIrank : the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Disc. Tod.* 21, 648–653. doi: 10.1016/j.drudis.2016.02.015

Chen, M., Vijay, V., Shi, Q., Liu, Z., and Fang, H. (2011). FDA-approved drug labeling for the study of drug-induced liver injury. *Drug Disc. Tod.* 16, 697–703. doi: 10.1016/j.drudis.2011.05.007

Franklin, J. M., Eddings, W., Austin, P. C., Stuart, E. A., and Schneeweiss, S. (2017). Comparing the performance of propensity score methods in healthcare database studies with rare outcomes. *Stat. Med.* 36, 1946–1963. doi: 10.1002/sim.7250

Harpaz, R., Dumouchel, W., Lependu, P. Bauer-Mehren A., Ryan, P., and Shah, N.H. (2013). Performance of pharmacovigilance signal-detection algorithms for the FDA adverse event reporting system. *Clin. Pharmacol. Therap.* 93, 539–546. doi: 10.1038/clpt.2013.24

Harpaz, R., Dumouchel, W., Shah, N. H., Madigan, D., Ryan, P., and Friedman, C. (2012). Novel data-mining methodologies for adverse drug event discovery and analysis. *Clin. Pharmacol. Therapeut.* 91, 1010–1021. doi: 10.1038/clpt.2012.50

Hastie, T., Tibshirani, R., and Friedman, J. (2009). "Boosting and additive trees," in *The Elements of Statistical Learning*, eds T. Hastie, R. Tibshirani and J. Friedman (New York, NY: Springer), 337–387.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statist. Med.* 29, 337–346. doi: 10.1002/sim.3782

Li, L., and Greene, T. (2013). A weighting analogue to pair matching in propensity score analysis. *Int. J. Biostatist.* 9, 215–234. doi: 10.1515/ijb-2012-0030

Marbac, M., Tubert-Bitter, P., and Sedki, M. (2016). Bayesian model selection in logistic regression for the detection of adverse drug reactions. *Biometr. J.* 58, 1376–1389. doi: 10.1002/bimj.201500098

Meinshausen, N., and Bühlmann, P. (2010). Stability selection. *J. R. Statist. Soc.* 72, 417–473. doi: 10.1111/j.1467-9868.2010.00740.x

Patorno, E., Grotta, A., Bellocco, R., and Schneeweiss, S. (2013). Propensity score methodology for confounding control in health care utilization databases. *Epidemiol. Biostatist. Public Health* 10:e8940-1. doi: 10.2427/8940

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55. doi: 10.1093/biomet/70.1.41

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., and Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* 20, 512–522. doi: 10.1097/EDE.0b013e3181a663cc

Seeger, J. D., Bykov, K., Bartels, D. B., Huybrechts, K., and Schneeweiss, S. (2017). Propensity score weighting compared to matching in a study of dabigatran and warfarin. *Drug Saf.* 40, 169–181. doi: 10.1007/s40264-016-0480-3

Seeger, J. D., Williams, P. L., and Walker, A. M. (2005). An application of propensity score matching using claims data. *Pharmacoepidemiol. Drug Saf.* 14, 465–476. doi: 10.1002/pds.1062

Stuart, E. A. (2010). Matching methods for causal inference : a review and a look forward. *Stat. Sci.* 25, 1–21. doi: 10.1214/09-STS313

Tatonetti, N. P., Ye, P. P., Altman, R. B., and Daneshjou, R. (2012). Data-driven prediction of drug effects and interactions. *Sci. Transl. Med.* 4:125ra31. doi: 10.1126/scitranslmed.3003377

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Statist. Soc.* 58, 267–288.

van Puijenbroek, E. P., Bate, A., Leufkens, H. G., Lindquist, M., Orre, R., and Egberts, A. C. (2002). A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. *Pharmacoepidemiol. Drug Saf.* 11, 3–10. doi: 10.1002/pds.668

Yoshida, K., Hernández-Díaz, S., Solomon, D. H., Jackson, J. W., Gagne, J. J., Glynn, R. J., et al. (2017). Matching weights to simultaneously compare three. *Epidemiology* 28, 387–395. doi: 10.1097/EDE.0000000000000627

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.