ORIGINAL RESEARCH ARTICLE

Use of Logistic Regression to Combine Two Causality Criteria for Signal Detection in Vaccine Spontaneous Report Data

Lionel Van Holle · Vincent Bauchau

Abstract

Purpose We evaluated the use of logistic regression to model the probabilities of spontaneously reported vaccine–event pairs being adverse reactions following immunization (ARFI), using disproportionality and unexpectedness of time-to-onset (TTO) distributions as predictive variables and the presence of events in the global product information as a dependent variable.

Methods We used spontaneous reports of adverse events from eight vaccines and their labels as proxies for ARFIs. Three logistic regressions were built to predict ARFIs based on different combinations of the proportional reporting ratio (PRR; disproportionality measure) and two Kolmogorov–Smirnov (KS) tests ('between vaccines' and the 'between events') of TTO distribution: model 1, using the PRR estimate and its 95 % lower confidence interval (CI) limit; model 2, using the *p* values of the two KS tests; and model 3, using the PRR (point estimate and lower CI limit) and both KS tests. The performance of the regressions (model fit statistics, calibration, and discrimination) was measured on 100 bootstrap samples.

Results Model 3, using two quantified causality criteria, provided the best performance for all measures. The *p* value of the 'between vaccines' KS test was the most significant predictive factor. Model 1 had the worst performance.

Conclusions Logistic regression allows estimation of the probability of a vaccine–event pair being an ARFI using two causality criteria at the population level assessed in spontaneous report data: the strength of association (disproportionality measure) and temporality (TTO distribution tests). Logistic regression combines and weights these causality criteria based on their respective ability to predict known safety issues.

Key points
The performance of three logistic regression models, incorporating different combinations of quantified causality criteria, was evaluated for the detection of safety signals from vaccine spontaneous report data
The logistic regression model integrating only the measure of the strength of association appeared to have the lowest performance for predicting known safety issues
The unexpectedness of the time-to-onset distribution for a given vaccine–event pair (when compared with the time-to-onset distribution of the same event reported following exposure to other vaccines) appeared to be best predictor of the reported event being a known safety issue
Logistic regression offers a framework in which quantified causality criteria can be combined to evaluate the probability of a vaccine–event pair being an adverse reaction following immunization based on our existing knowledge of vaccine safety profiles

L. Van Holle (✉) · V. Bauchau
Vaccine Safety Research Group (VSRG), Vaccines Clinical Safety and Pharmacovigilance (VCSP), GlaxoSmithKline Vaccines, Parc de la Noire Epine, Avenue Fleming, 20, 1300 Wavre, Belgium
e-mail: lionel.f.van-holle@gsk.com

1 Introduction

Data mining algorithms (DMAs) have been developed for screening spontaneous report databases (SRDs). The majority of these algorithms detect product–event pairs (P–Es) presenting a disproportionate number of reports compared with the expected number from other/all products and other/all events within the same SRD [1, 2]. This measure of disproportionality offers a proxy of the strength of association between a product and an event while accounting for the absence of exposure data characteristics of spontaneous data [3].

These DMAs were first developed for screening the SRDs held by regulatory authorities: the Empirical Bayes Geometric Mean (EBGM) for the Food and Drug Administration SRD [2], the information component (IC) for the World Health Organization (WHO) SRD, the proportional reporting ratio (PRR) for the UK SRD, and the reporting odds ratio for the Netherlands Pharmacovigilance Foundation Lareb SRD. Over time, the use of these DMAs extended to SRDs held by drug and vaccine manufacturers. In this study, we focus on the GlaxoSmithKline (GSK) vaccines SRD containing spontaneously reported adverse events (AEs) following immunization by a GSK vaccine.

These DMAs all share the same objective: to estimate the strength of association. However, this is only one of a number of causality criteria at the population level for determining whether a vaccine may have caused a particular AE (others include temporal relationship, dose-response relationship, consistency of evidence, specificity, biological plausibility, and coherence) [4]. The use of the causality criterion strength of association does not necessarily require prior medical insight or external data sources. DMAs have thus focused only on strength of association, allowing signals of disproportionate reporting to be generated autonomously and in an automated way for all P–Es.

According to the WHO, establishment of the temporal relationship as a causality criterion at the population level is based on the principle that, "vaccine exposure must precede the occurrence of the event" [4]. With a few exceptions, this is mostly the case for events reported in SRDs, whether causally or just coincidentally related to vaccination. We recently demonstrated that a temporal relationship for a vaccine–event (V–E) pair from an SRD could be quantified by measuring the deviation of its time-to-onset (TTO) distribution from the overall patterns of reported TTO of that vaccine with other reported AEs and of that AE with other vaccines [5–7]. In other words, temporality could be quantified by measuring the unexpectedness of the reported TTO distribution, just as the strength of association is quantified by measuring how unexpected the number of reports is for a given V–E pair.

This allowed the development of a new generation of DMAs able to screen SRDs to flag P–Es with unexpected TTO distributions autonomously and automatically, without prior medical insight or other data sources.

As stressed in the original proof-of-concept study [5], the two types of DMAs—TTO and disproportionality (strength of association)—are complementary theoretically and in their limitations. The TTO DMA is based on TTO data, which are neglected by the disproportionality DMA and recognised to be an important criterion to assess possible causality during medical evaluation of individual case reports. On the other hand, TTO DMA can only be performed on the subset of spontaneous reports presenting TTO values within the window of interest. It excludes spontaneous reports for which TTO information is missing or occurs after the predefined time window. Consequently, TTO DMAs may miss the detection of P–E pairs that have only a small number of reports with available TTO information. Disproportionality DMAs require adjustment to account for different reporting rates between demographic or secular strata, but can be performed on uncommon or long-term AEs.

The use of TTO DMAs raises the practical problem of quantitative signals that can be generated by either unexpected numbers of reports or unexpected TTO distributions. The flagging of P–E pairs as quantitative signals only when they are detected as both disproportionate and temporal signals would result in a signal detection system with lower sensitivity and higher specificity than either individual method alone. Knowing that we would systematically lose the ability to detect uncommon and long-term events, this option is not viable for a signal detection system. On the other hand, flagging P–E pairs that are unexpected either in terms of number of reports or in TTO distribution would result in a signal detection system with low specificity and high sensitivity [6]. Consequently, further methodological research was needed to build a signal detection algorithm that could account for two, and potentially more, quantified causality criteria at the population level.

The logistic regression framework was selected and analysed for its potential to combine multiple factors, and because previous papers have demonstrated the usage of logistic regression to weight causality criteria at the individual level to model medical expert judgement [8, 9].

Here, we illustrate how logistic regression can be used to model the probability of a V–E pair being an ARFI, meaning an AE causally related to immunization, using disproportionality and unexpectedness of the TTO distribution as predictive variables and the presence of events in the global product information (GPI) as a predicted dependent variable. The estimated parameters of the logistic regression provide the weight of each causality

criterion to define the probability of being an ARFI [10], using the current knowledge of the V–Es already recognized as being a safety concern, a piece of information neglected by both disproportionality and TTO methods. We use this approach for the two causality criteria at the population level that can currently be automatically and autonomously assessed with DMAs from the SRD without prior medical knowledge.

2 Methods

2.1 The Proportional Reporting Ratio

We selected the PRR [1, 11] for the disproportionality measure as we highlighted that measures based on the relative reporting ratio, like the EBGM or IC, are biased downwards when used on the GSK Vaccines SRD [12].

The PRR is calculated based on a 2×2 table, as in Table 1:

The PRR can be expressed as

$$\text{PRR} = \frac{A/(A\ +\ B)}{C/(C+D)} = \frac{A \times (C+D)}{C \times (A+B)}$$

where A, B, C, and D are defined in Table 1. A 95 % confidence interval around the PRR can be derived [1]:

$$e^{\ln(\text{PRR}) \pm \sqrt{\left(\frac{1}{A} - \frac{1}{A+B} + \frac{1}{C} - \frac{1}{C+D}\right)}}$$

To account for demographic and secular differences between vaccinated populations, the PRR was stratified by sex, age, region, and year by using a Mantel–Haenszel measure of effect [13].

We considered the stratified PRR estimate (PRR_E) to summarize the strength of association, and its 95 % lower confidence limit (PRR_{LL}) to account for measure variability, as both measures are often used in DMA based on PRR [14].

2.2 Time-to-Onset Signal Detection

TTO signal detection is a non-parametric DMA for detecting V–Es with a TTO distribution that is significantly different from:

– the TTO distribution of the same vaccine with the other reported events ('between events' test)

And

– the TTO distribution of the same event reported after administration of other vaccines ('between vaccines' test)

at a given significance alpha level and within a given time window [5]. The two-sample Kolmogorov–Smirnov (KS) test statistic is sensitive to differences in the distribution from which the two samples were drawn, such as differences in location, dispersion, or skewness.

Here, we use the two p values generated by the 'between events' and 'between vaccines' KS tests to summarize the unexpectedness of TTO data over the 60-day period after vaccination. The time window of 60 days was previously associated with high performance in terms of positive predictive value [6].

The algorithm identifies an unexpected TTO distribution for a V–E through detection of TTO distributions that deviate from the overall reported TTO distributions for other reported events with the vaccine of interest and for the event of interest with other vaccines. The assumption that underpins this approach is that most reported V–E pairs are not causally related, so that the overall TTO distributions are dominated by reporting biases and noise [7]. This assumption that most reported V–E pairs are not causally related also underpins the disproportionality approach and, if violated, generates the so-called masking effect [15, 16].

2.3 Data Selection

For practicality reasons, the calculation of PRR estimates and KS p values was restricted to eight vaccines: Rotarix™, Engerix™, Cervarix™, Fluarix™, Infanrix™, Infanrix™ Hib, Havrix™, and Twinrix™. These vaccines together represented more than half of the vaccine spontaneous reports in the SRD at the data lock point date of 1 February 2010 and covered a diverse range of vaccine characteristics. They were thus considered representative of the entire SRD at GSK vaccines (Tables 2, 3). The entire SRD was used to compute the PRR and KS p values for these eight vaccines.

2.4 The Dependent Variable

The dependent variable ('ARFI') was based on the safety information from the GPI of each vaccine.

For each V–E, the Medical Dictionary for Regulatory Activities (MedDRA)[1] Preferred Terms corresponding to a medical term listed in the GPI for that vaccine were assigned

Table 1 Contingency table

	Reports with the event of interest	Reports without the event of interest
Reports with the vaccine of interest	A	B
All other reports	C	D

Table 2 Description of the therapeutic indication of the vaccines under study

Vaccine	Therapeutic indication (extracted from http://www.medicines.org.uk/emc on 14 August 2014)
Engerix™	Active immunization against hepatitis B virus infection caused by all known subtypes in non-immune subjects
Havrix™ (adult and pediatric)	Active immunization against infections caused by hepatitis A virus
Cervarix™	Vaccine for use from the age of 9 years for the prevention of premalignant genital (cervical, vulvar, and vaginal) lesions and cervical cancer causally related to certain oncogenic human papillomavirus types
Infanrix™	Vaccine indicated for booster vaccination against diphtheria, tetanus, pertussis, and poliomyelitis diseases in individuals from 16 months to 13 years of age inclusive who have previously received primary immunization series against these diseases
Infanrix™ Hib	Active immunization against diphtheria, tetanus, pertussis, poliomyelitis and *Haemophilus influenzae* type b disease from the age of 2 months
Rotarix™	Active immunization of infants aged 6–24 weeks for prevention of gastroenteritis due to rotavirus infection
Fluarix™	Prophylaxis of influenza, especially those who run an increased risk of associated complications. Fluarix™ is indicated in adults and children from 6 months of age
Twinrix™ (adult and pediatric)	Indicated for individuals who are at risk of both hepatitis A and hepatitis B infection

the value 1 and the others the value 0. The list of events in the GPI is considered as a proxy of the list of events causally related to the vaccine. Indeed, medical terms in the GPI are generated from either clinical or post-marketing experience. For data obtained from randomized clinical trials, a significant excess of cases in the vaccine group compared with a control can be causally attributed to the vaccine at a given significance level due to the properties of randomized clinical trials. Post-marketing data may be generated from a variety of settings, such as pharmaco-epidemiological studies, electronic health records, and spontaneous reports; when there is no equivalent of a randomized study, potential signals may be highly biased and are consequently usually subject to evaluation based on causality criteria at the population and individual levels [4] before being included in the GPI. However, not all medical terms followed this process before being included in the GPI. In addition, listed medical terms had to be mapped to MedDRA preferred terms for consistency with spontaneous report data, which are coded using the MedDRA dictionary. Consequently, the ARFIs used could be considered as mainly, if not completely, based on causality assessments.

2.5 Logistic Regression Models

Logistic regression models the relationship between a dependent binary variable (the 'ARFI' in this case) and predictive variables. For any V–E pair, an estimated probability of being an ARFI can be derived based on the estimated model parameters.

Three different models, characterized by different choices of predictive variables, have been studied:

Model 1 Using disproportionality information only

$$\mathrm{logit}(\mathrm{ARFI}(1)) = \boldsymbol{\alpha}^1 + \boldsymbol{\beta}_1^1 \mathrm{PRR}_E + \boldsymbol{\beta}_2^1 \mathrm{PRR}_{LL}.$$

The logistic regression modelled the probability of a V–E being an ARFI based on the disproportionality measure: the stratified PRR and its 95 % lower limit.

These two predictive variables may have missing values, for example in the case of a vaccine causing a rare event, which would then be likely to be reported solely after the vaccine of interest and never with other vaccines. As missing values cannot be handled as such by the logistic regression model, it was important to categorize the PRR_E and its PRR_{LL}. The two variables were thus categorized as follows[2]: 'N/A'; '[0, 0.8]'; ']0.8, 1.2]'; ']1.2, 2]'; ']2, 5]'; ']5, 10]'; ']10, 100]'; '100+'.

Model 2 Using unexpectedness of the TTO distribution only

$$\mathrm{logit}(\mathrm{ARFI}(1)) = \boldsymbol{\alpha}^2 + \boldsymbol{\beta}_1^2 \mathrm{KS}_{BE} + \boldsymbol{\beta}_2^2 \mathrm{KS}_{BV}.$$

The logistic regression modelled the probability of a V–E being an ARFI based on the unexpectedness of the TTO distribution, summarized by the *p* value of the 'between events' (KS_{BE}) and 'between vaccines' (KS_{BV}) KS tests.

The *p* values KS_{BE} and KS_{BV} were categorized as follows: 'N/A'; '0'; '[Min, Q1['; '[Q1, Median['; '[Median, Q3['; '[Q3, 0.01]'; ']0.01,1]', where Min, Q1, Median, and Q3 correspond to the minimum, first quartile, median, and third quartile observed in the interval]0, 0.01] for the *p* values KS_{BE} and KS_{BV}, respectively. This dynamic categorization should ensure interpretability and that each category contains a sufficient number of observations.

Model 3 Using both the disproportionality and the unexpectedness of the TTO distribution.

---

2 [A,B] refers to an interval between A and B, both values included in the interval whereas]A,B,[refers to an interval between A and B, both values excluded.

Table 3 Characteristics of spontaneous reports in the GlaxoSmithKline Vaccines spontaneous report database, by vaccine

Vaccine	Age at event (years): median (Q1–Q3)	Female (%)	Year of reporting: median (Q1, Q3)	Number (%) of spontaneous reports	Number of countries
Engerix ^{TM}	31.0 (18.0, 43.0)	64.2	1999 (1993, 2005)	34,347 (23.4 %)	92
Havrix ^{TM}	23.0 (11.0, 40.0)	57.8	2004 (1998, 2007)	9,066 (6.2 %)	58
Cervarix ^{TM}	15.0 (12.0, 17.0)	99.5	2009 (2008, 2009)	3,437 (2.3 %)	63
Infanrix ^{TM}	5.0 (1.5, 10.0)	45.5	2006 (2003, 2007)	9,732 (6.6 %)	59
Infanrix ^{TM} Hib	1.5 (0.8, 1.9)	42.5	2002 (1999, 2003)	1,027 (0.7 %)	21
Rotarix ^{TM}	0.3 (0.2, 0.6)	46.3	2008 (2007, 2009)	2,800 (1.9 %)	73
Fluarix ^{TM}	41.0 (19.0, 60.0)	60.0	2005 (2002, 2007)	6,864 (4.7 %)	69
Twinrix ^{TM}	31.0 (19.0, 45.0)	57.6	2006 (2003, 2008)	9,836 (6.7 %)	51

The logistic regression modelled the probability of a V–E being an ARFI based on the disproportionality measure and the unexpectedness of the TTO distribution.

$$\text{logit}(\text{ARFI}(1)) = \alpha^3 + \beta_1^3 \text{PRR}_E + \beta_2^3 \text{PRR}_{LL} + \beta_3^3 \text{KS}_{BE} + \beta_4^3 \text{KS}_{BV}$$

with the same categorization as for model 1 and 2.

2.6 Measures of Performance

The performance of a logistic regression can be summarized by the following characteristics:

Model fit statistics A global test (Wald test) measures how likely it is that the group of predictive variables could be of no use in predicting the value of the dependent variable ('ARFI' here). The more unlikely (small *p* values), the better the model fits the data [17].
Discrimination The concordance statistic (also known as C statistic or area under the curve) [18] measures the probability that a random listed V–E pair has a higher probability than a random non-listed V–E pair. The closer to 1, the better the model discriminates.
Calibration This refers to the agreement between the observed and predicted outcome for the dependent variable ('ARFI' here). The widely used Hosmer–Lemeshow test [19] tests the null hypothesis that there is no difference between the observed and predicted values of the response variable. The more unlikely (small *p* values), the worse the calibration.

Steyerberg [20] showed that bootstrap resulted in the most accurate estimate of model performance, providing a bias close to zero. Bootstrapping replicates the process of sample generation from an underlying population, of the same size as the original data set, by drawing samples with replacement from the original data set. We consequently took 100 bootstrap repetitions of the entire GSK Vaccines SRD and, for each one, performed the KS tests, calculated the PRR, and ran the three logistic regression models described above. For each bootstrap repetition and each

logistic regression, we measured the different performance criteria of the logistic regression model applied to the subset of eight vaccines described above.

The performance of each of these models was described graphically with box plots showing the distribution of the median and first and third quartile values (indicated by the middle, top, and bottom lines of the box, respectively). The interquartile range, containing the middle 50 % of the data, is thus represented by the vertical length of the box, whilst the range of the data is the vertical distance between the smallest and largest values, including or excluding outliers.

The impact of the predictive variables categories on the estimated probability values was evaluated. The estimated probability distribution was also compared between the sources of the data included in the GPI (clinical development or post-marketing experience).

The results and figures were produced using SAS9.2. The following procedures were used: PROC NPAR1WAY for the calculation of the two-sample KS test *p* values and PROC LOGISTIC for the logistic regression.

3 Results

The original dataset contained 9474 V–Es to be modelled for their probability of being an ARFI, using the three logistic regression models based on data from the eight vaccines under study; 803 (8.5 %) were considered as ARFIs based on the safety information from the GPI. Over the 100 bootstrap samples, there were an average of 7,831 different V–Es, of which 9.2 % on average were considered as ARFIs.

3.1 Model Fit Statistics

The global Wald test showed that the three logistic models were highly significant. The two most significant models were model 2 (using only the KS test *p* values) and model 3 (using the KS test *p* values and the PRR), followed by model 1 using only the PRR (Fig. 1).
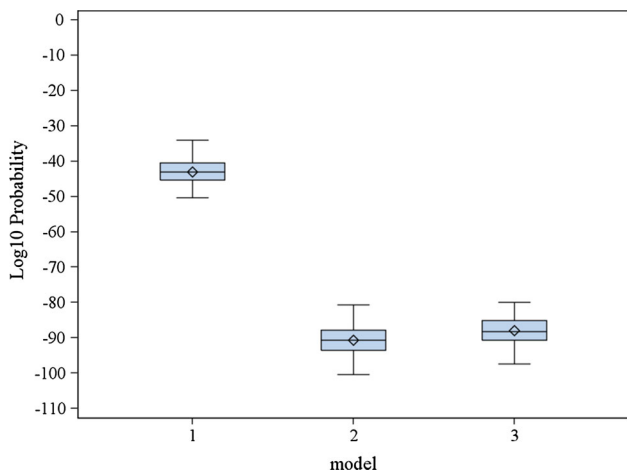
Fig. 1 Wald test *p* value distribution for the test of the null hypothesis that beta = 0 for the logistic regression models 1, 2, and 3

For model 1, both the PRR_E and the PRR_{LL} were highly significant predictive variables at similar alpha levels. For model 2, a considerable difference in significance was highlighted between KS_{BV} (highly significant) and KS_{BE} (not significant at alpha level = 0.01) predictive variables. For model 3, KS_{BV} was the most significant predictive variable, followed by the PRR_E. The PRR_{LL} factor was borderline, with a significance level of 0.01; the KS_{BE} factor was not significant.

3.2 Discrimination

Model 3 discriminates between the GPI-listed and unlisted V–Es better than do models 1 and 2 (Fig. 2).

3.3 Calibration

The distribution of the *p* values for the Hosmer–Lemeshow test shows that the null hypothesis (no difference between
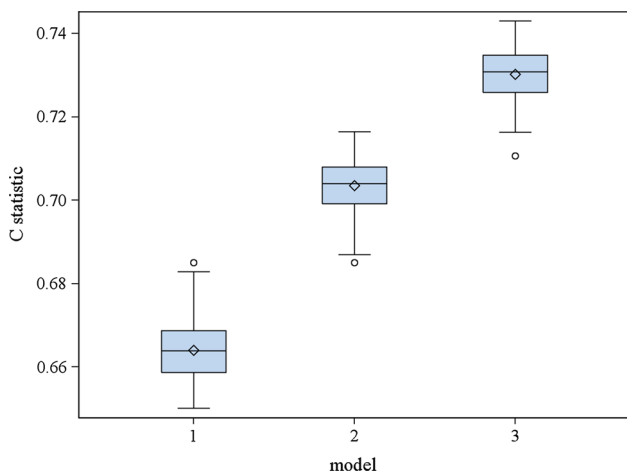
Fig. 2 Area under the receiver operating curve (C statistic) distribution for the three logistic regression models
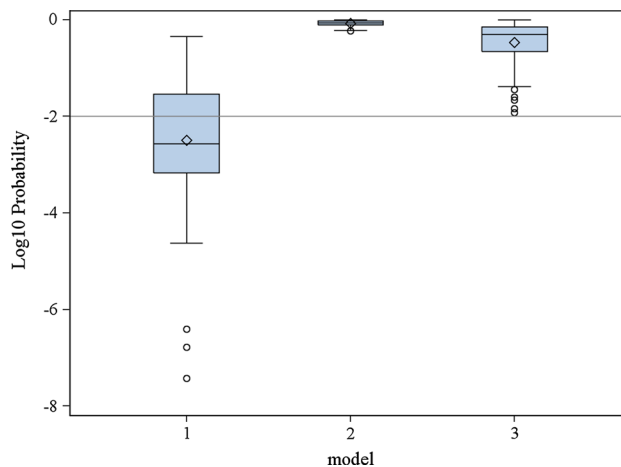
Fig. 3 Hosmer–Lemeshow test *p* value distribution for the three logistic regression models

observed and predicted values) was not rejected at alpha level 0.01 (represented by a horizontal line across the graph) for any bootstrap samples used for logistic regression models 2 and 3 (Fig. 3). However, the null hypothesis was rejected for 61 of 100 bootstrap samples for model 1. This suggests that the logistic regression model was well calibrated when the *p* values of the KS tests were used as predictive variables (as in models 2 and 3) but not when only the stratified PRR and its lower limit were used as predictive variables (as in model 1).
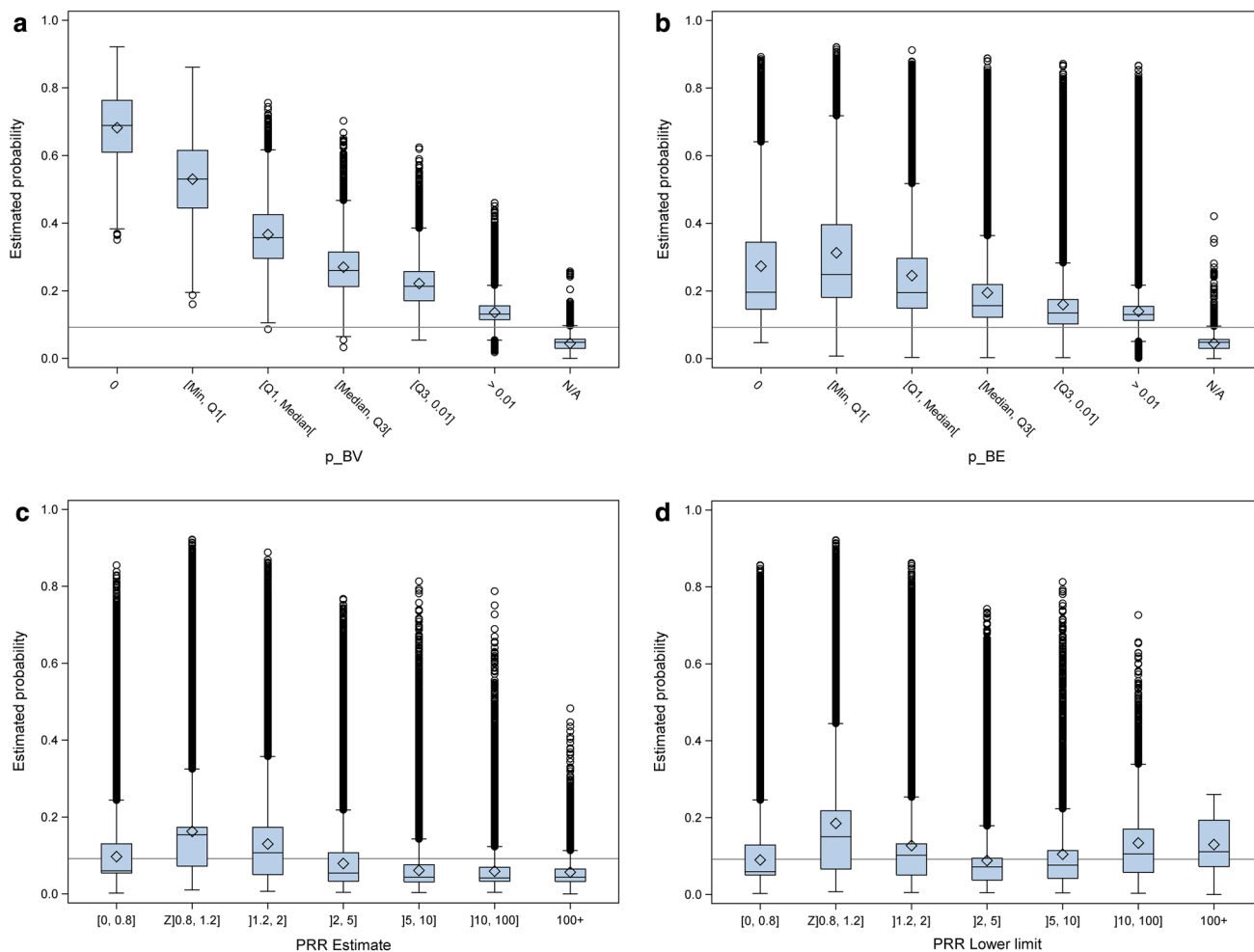
3.4 Distribution of the Estimated Probability

Figure 4 shows the monotonic relationship between the *p* value KS_{BV} and the estimated probability of a V–E being an ARFI by the model 3: the lower the *p* value, the higher the estimated probability. V–E with very low KS_{BV} *p* values (0 or in the first quartile of values in the interval]0, 0.01]) have an estimated probability far above the average percentage of listed V–Es. For example, V–Es presenting a null KS_{BV} *p* value have a median probability around 70 % (Fig. 4—upper left panel).

The KS_{BE} *p* value does not show such a monotonic relationship with the estimated probability. The category with the highest median estimated probability has an estimated probability around 20 % only (Fig. 4—upper right panel).

The relationship between the PRR estimate (lower limit) and the estimated probability is nonlinear, with a local maximum in the median estimated probability for the ']0.8, 1.2]' (']0.8, 1.2]') category followed by a local minimum for the ']10, 100]' (']0, 0.8]') category.

The median estimated probability of listed V–Es was the same whatever the source: clinical development or post-marketing (Fig. 5). However, the mean estimated

Fig. 4 Distribution of probability estimated by model 3 for each category of the different parameters: **a** P_{BV}, **b** P_{BE}, **c** PRR_{LL}, and **d** PRR_E. The *horizontal line* represents the average percentage of vaccine–event pairs listed in the global product information. *BE* between events, *BV* between vaccines, *E* estimate, *LL* lower limit, *PRR* proportional reporting ratio
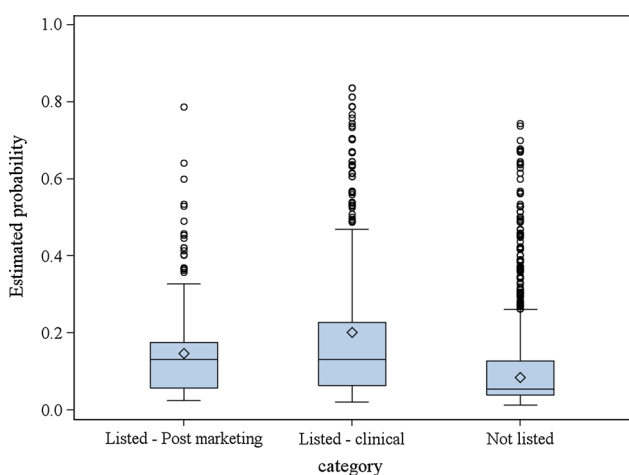
Fig. 5 Distribution of the estimated probability according to the source of data having led to some events to be listed in the global product information

probability was higher for ARFIs detected at the clinical level. This could be because some of these ARFIs may present a very distinctive pattern in terms of disproportionality and TTO distribution. Regardless of the data source, the estimated probability was higher for ARFIs than for the not listed events.

As an example, model 3 gave the highest probability of being an ARFI for the ten V–E pairs shown in Table 4.

None of these V–E pairs would have been detected by the stratified PRR when using a threshold of two on the 95 % lower limit, except for the pair Rotarix[TM]–Diarrhoea. However, a TTO signal would have been generated for all of them, except for the pair Twinrix[TM]–Fatigue, using a threshold of 0.01 for the *p* value of both KS tests.

Model 1, which uses only disproportionality information, estimates a higher probability (36 %) for V–Es having $PRR_E = PRR_{LL} = '$]$0.8, 1.2$]$'$ because it is within this range

Table 4 Ten vaccine–event pairs for which model 3 gave the highest probability of being an adverse reaction following immunization

Vaccine: event	Listed?	PRR_E	PRR_{LL}	KS_{BE}	KS_{BV}	Prob model 1 (%)	Prob model 2 (%)	Prob model 3 (%)
Engerix™: Myalgia	Yes	]0.8, 1.2]	]0.8, 1.2]	[Min, Q1[	0	36	84	93
Infanrix™: Pyrexia	Yes	[0, 0.8]	[0, 0.8]	[Min, Q1[	0	9	84	86
Rotarix™: Diarrhoea	Yes	]10, 100]	]10, 100]	[Min, Q1[	[Min, Q1[	16	86	84
Engerix™: Pruritus	Yes	]0.8, 1.2]	]0.8, 1.2]	0	[Min, Q1[	36	69	83
Engerix™: Vomiting	Yes	]0.8, 1.2]	]0.8, 1.2]	[Min, Q1[	[Q1, Median[	36	70	83
Engerix™: Abdominal pain	Yes	]1.2, 2]	]0.8, 1.2]	[Q1, Median[	[Min, Q1[	24	74	82
Twinrix™: Fatigue	Yes	]0.8, 1.2]	]0.8, 1.2]	>0.01	[Min, Q1[	36	67	82
Engerix™: Arthralgia	Yes	]0.8, 1.2]	]0.8, 1.2]	0	0	36	65	82
Havrix™: Headache	Yes	]0.8, 1.2]	[0, 0.8]	[Median, Q3[	0	12	77	81
Cervarix™: Pyrexia	Yes	]0.8, 1.2]	[0, 0.8]	[Median, Q3[	0	12	77	81

BE between events, *BV* between vaccines, *E* estimate, *KS* Kolmogorov–Smirnov, *LL* lower limit, *Prob* estimated probability, *PRR* proportional reporting ratio

of values that the observed frequency of known safety issues was observed. Models 2 and 3 estimate a higher probability for V–Es with small *p* values for the KS tests, and model 3 fluctuates around these probabilities to take into account the disproportionality information. When $PRR_E = PRR_{LL} = $ ']0.8,1.2]', model 3 estimates higher probabilities than does model 2.

4 Discussion

Our analyses have shown that the logistic regression can be used to predict ARFI based on the combination of several predictive causality criteria at the population level. Among the combinations tested, the logistic regression based both on KS *p* values and on PRR provided the best model in terms of fit, calibration, and discrimination. The logistic regression model based on KS *p* values only (model 2) provided similar performance results in terms of fit and calibration but lower performance in terms of discrimination. The logistic regression model based solely on PRR (model 1) gave the poorest performance for all measures.

In model 1, the disproportionality information summarized by the PRR estimate and its 95 % lower limit poorly predicted the presence of AEs in the GPI for the eight vaccines under study. The unexpectedness of a TTO distribution, used in model 2 and 3, was a better predictor of the presence of AEs in the GPI than the disproportionality information used in model 1.

Taking the GPI as a proxy of the list of events causally associated with the vaccines, we can conclude that temporality seems to be a stronger predictor of causality than the strength of association for the eight vaccines under study, at least when temporality and strength of association are estimated in the context of spontaneous report data.

This highlights the importance of using this quantified and objective temporality criterion for signal detection in the SRD. More specifically, the more confidently one can reject, for a specific event, the null hypothesis of a common TTO distribution between the vaccine of interest and the other vaccines (KS_{BV}), the higher the estimated probability of a causal association between that event and the vaccine of interest. On the other hand, the *p* value of the KS_{BE} was evaluated by both models 2 and 3 as not being a significant predictive factor of causality, at least when used with KS_{BV}. The diverse categories of AEs may generate differences in the reported TTO distribution independently from causal association between the vaccine and event.

Logistic regression has several advantages for improving quantitative signal detection. First, it uses current knowledge of the safety profile of the vaccines under post-marketing pharmacovigilance for attributing weights to the different measures of unexpectedness, in terms of number of spontaneous reports and TTO distribution. The model can be calibrated on the actual SRD of interest and does not need predefined thresholds extrapolated from other SRDs with different characteristics or from occasional retrospective performance evaluations.

Second, the logistic regression model allows the linear combination of predictive factors of causality. Causality assessment is driven by several complementary criteria. The fact that logistic regression can combine the use of two causality criteria at the population level (the strength of association and a more refined notion of temporality) provides an elegant solution for coping with the complementarity of these two measures, as previously highlighted [5].

Third, logistic regression solves the dilemma of what threshold to use for defining disproportionate signals. The current practice in quantitative signal detection is to treat disproportionality scores dichotomously: above a given

threshold there is a quantitative signal and below it there is no signal. We previously showed that published recommendations on the use of thresholds may not be optimal [12] depending on the SRD characteristics. The determination of the 'ideal' threshold is complex and crucial in terms of signal detection performance. By using categorized values of the different measures of unexpectedness, we overcome the uncertainty surrounding the 'best' threshold to use. Indeed, the logistic regression model automatically attributes higher weights to the categories with the highest predictive value, based on the current knowledge of the safety profile. It reduces the dependence to the choice of a unique threshold (even if they are still dependent on our choices of categories). Some events are solely reported after a given immunization, not because they are caused by the vaccination, but sometimes because the report is about a lack of efficacy of the vaccine. For example, the AEs 'Rotavirus infection' or 'Rotavirus test positive' are unlikely to be spontaneously reported after any vaccination other than Rotarix^{TM}. Consequently, these two events are characterized by very high values of PRR_{LL}. They actually fall under the category ']10, 100]'. Depending on how frequently an event listed in the GPI was characterized by a PRR_{LL} in the category ']10, 100]', the logistic regression weights this category for predicting ARFIs.

Fourth, logistic regression based on strength of association and temporality can provide a score reflecting the probability of a V–E being an ARFI. This is an intuitive score for physicians and other non-statisticians. It can be used directly as a signal detection algorithm: V–Es flagged with a high probability of being an ARFI (based on strength of association and temporality) and not yet in the GPI may present the highest probability of a causal association between the vaccine and the event or at least share characteristics of events already listed in the GPI. However, using a logistic regression model directly as a signal detection algorithm brings challenges that will need careful prospective evaluation. Indeed, including more causality criteria in the logistic regression lowered our ability to detect signals when the KS_{BV} was missing. Indeed, when KS_{BV} is missing, the estimated probability based on the other predictive variables (KS_{BE}, PRR_E, and PRR_{LL}) will always be low, as these variables are poor predictors. The inclusion of several causality criteria in a signal detection system partially replicates, at an aggregate level, the process of signal evaluation where insufficient information may prevent a conclusion from being drawn.

A hidden assumption behind our logistic regression model is that the safety profile of the vaccine is for the most part known and summarized in the GPI given the pre-marketing data from clinical trials and parallel methods for detecting signals including literature reviews, post-authorization safety studies, and medical reviewing. Otherwise,

the logistic regression would be fitted based on too high a proportion of V–Es being misclassified as not causally associated, which could reduce the model performance for detecting ARFIs. Furthermore, defining the dependent variable as the presence of the event in the GPI makes the 'ARFI' a time-evolving dependent variable. A dependent variable reflecting live changes in the GPI could generate instability in the estimation of the parameter, leading to instability in the estimated probability of V–Es being ARFIs. Additional prospective research should be conducted to monitor the stability of the predicted probabilities over time. The other assumption underlying these logistic regression models is that the measures of unexpectedness that are most strongly associated with known safety problems are those that will also allow us to detect as yet unknown safety problems.

Previous observations [6] suggest that the detection of signals based on unexpected TTO distributions requires a larger number of case reports than the detection of signals based on disproportionate reporting, since the cases with missing TTO information cannot be used by KS_{BV} and KS_{BE}. Consequently, the use of logistic regression could delay signal detection, at least for signals that had the potential to be detected by their disproportionality profile alone. On the other hand, the use of the aggregate and weighted information about unexpected TTO distribution and strength of association may flag new V–Es worth further evaluation.

Finally, logistic regression offers a framework allowing the use of several causality criteria along with current knowledge of the safety profiles under monitoring. Additional research should be conducted to quantify the other causality criteria at the population level, beyond 'strength of association' and 'temporality'. For example, 'specificity' could be captured as the percentage of reports for which the vaccine was the only plausible cause for explaining the AE post-immunization (or the 95 % binomial lower limit of that percentage to account for variability). The 'consistency of evidence' causality criteria could be a measure of concordance between what has been measured in the SRD under monitoring and another source (such as registries or observational data). If the logistic regression models integrating these additional causality criteria appear to perform better than the one with temporality and strength of association, only then should we consider incorporating these new quantified causality criteria.

In this study, the theoretical and practical relevance of the logistic regression framework was analysed on vaccine spontaneous report data. However, we envision this framework to be also applicable to drugs, other SRDs, and observational electronic healthcare databases. Different settings may be needed to take into account specificities of the products and database holders, and the dependent

variable can be defined differently to facilitate early detection. We take as reference the recent research paper from Caster [21], where a shrinkage logistic regression model was applied on Vigibase spontaneous report data to model the probability that a drug–event pair is an emergent safety signal. Instead of using solely the causality factors as potential predictors of being an emergent safety signal, they pragmatically used the different aspects of strength of evidence based on report quality and content. A measure of the unexpectedness of TTO distribution (originally developed for vaccine spontaneous reports and not yet assessed on drug spontaneous reports) was not used by the model but only a crude estimate of the plausibility of the reported TTO. The logistic regression framework could easily integrate this refined notion of temporality and would automatically weight it relative to the other aspects of strength of evidence. Indirectly, it would also assess if it is as good a predictor for drug emerging safety signals as it was for events listed in the GPI of the GSK vaccines under study.

5 Conclusion

The logistic regression framework allows the combined use of two causality criteria—the strength of association (estimated by a disproportionality measure) and the temporality (estimated by a KS test)—to estimate from spontaneous report data the probability that a V–E pair is an ARFI. Logistic regression optimally weights the causality criteria and combines them based on their ability to predict known safety issues. A prospective evaluation of this method is needed to evaluate its potential added value in the pharmacovigilance toolkit.

References

1. Van Puijenbroek EP. A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol Drug Saf. 2002;11:3–10.
2. Dumouchel W. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. Am Stat. 1999;53:177–90.
3. Goldman SA. Limitations and strengths of spontaneous reports data. Clin Ther. 1998;20:C40–4.
4. World Health Organization, Causality assessment of an adverse event following immunization (AEFI)—user manual for the revised WHO classification, WHO/HIS/EMP/QSS. 2013. http://www.who.int/vaccine_safety/publications/aevi_manual.pdf. Accessed 7 July 2014.
5. Van Holle L. Using time-to-onset for detecting safety signals in spontaneous reports of adverse events following immunization: a proof of concept study. Pharmacoepidemiol Drug Saf. 2012;21:603–10.
6. Van Holle L. Signal detection on spontaneous reports: a comparison of the performance of a method based on disproportionality and a method based on the time from immunization to onset of adverse events. Pharmacoepidemiol Drug Saf. 2014;23:178–85.
7. Karimi G. Time-to-onset in spontaneous reports: the possibility to detect the unexpected. Pharmacoepidemiol Drug Saf. 2013;22:556–7.
8. Arimone Y. A new method for assessing drug causation provided agreement with experts' judgement. J Clin Epidemiol. 2006;59:308–14.
9. Théophile H. An updated method improved the assessment of adverse drug reaction in routine pharmacovigilance. J Clin Epidemiol. 2012;65:1069–77.
10. DeCarlo LT. Signal detection theory and generalized linear models. Psychol Methods. 1998;3(2):186–205.
11. Evans SJW. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf. 2001;10:483–6.
12. Van Holle L. The upper bound to the relative reporting ratio—a measure of the impact of the violation of hidden assumptions underlying some disproportionality methods used in signal detection. Pharmacoepidemiol Drug Saf. 2014;23:787–94.

13. dos Santos Silva I. Cancer epidemiology: principles and methods. Lyon: International agency for research on cancer; 1999. p. 309–11.

14. Deshpande G. Data mining in drug safety—Review of published threshold criteria for defining signals of disproportionate reporting. Pharm Med. 2010;24(1):37–43.

15. Gould AL. Practical pharmacovigilance analysis strategies. Pharmacoepidemiol Drug Saf. 2003;12:559–74.

16. Maignen F. A conceptual approach to the masking effect of measures of disproportionality. Pharmacoepidemiol Drug Saf. 2014;23:208–17.

17. Kleinbaum DG. Logistic regression: a self-learning text. 2nd ed. Berlin: Springer; 2002. p. 130–6.

18. Michael EM. Validation of probabilistic predictions. Med Decis Making. 1993;13:49–57.

19. Hosmer DW Jr, Lemeshow S. Applied logistic regression. 2nd ed. New York: Wiley; 2000.

20. Steyerberg EW. Internal validation of predictive models—efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54:774–81.

21. Caster O. Improved statistical signal detection in pharmacovigilance by combining multiple strength-of-evidence aspects in vigiRank—retrospective evaluation against emerging safety signals. Drug Saf. 2014;37:617–28.