TEDAR: Temporal dynamic signal detection of adverse reactions

Antonino Aparo[a], Pietro Sala[a,*], Vincenzo Bonnici[a], Rosalba Giugno[a,*]

[a] *Department of Computer Science, University of Verona, 37134 Strada le Grazie 15, Verona, Italy.*

Abstract

Computational approaches to detect the signals of adverse drug reactions are powerful tools to monitor the unattended effects that users experience and report, also preventing death and serious injury. They apply statistical indices to affirm the validity of adverse reactions reported by users. The methodologies that scan fixed duration intervals in the lifetime of drugs are among the most used. Here we present a method, called TEDAR, in which ranges of varying length are taken into account. TEDAR has the advantage to detect a greater number of true signals without significantly increasing the number of false positives, which are a major concern for this type of tools. Furthermore, early detection of signals is a key feature of methods to prevent the safety of the population. The results show that TEDAR detects adverse reactions many months earlier than methodologies based on a fixed interval length.

Keywords: `elsarticle.cls`, LaTeX, Elsevier, template

2010 MSC: 00-01, 99-00

*Corresponding authors

Email addresses: `pietro.sala@univr.it` (Pietro Sala), `rosalba.giugno@univr.it` (Rosalba Giugno)

1. Introduction

Adverse drug reactions (adr) are unsuspected and unpleasant reactions emerging from the use of medical products, in particular drugs. The safety of drugs is still a major public health concern that causes many deaths and severe injuries[1]. National programs collect spontaneous reports in which associations between drugs and adrs are observed by citizens or doctors. Reports are stored in surveillance databases[2, 3, 4]. They cover a variety of user categories, drugs and effects of such drugs over the time the drugs are on the market (i.e. their lifetime). In the context of pharmacovigilance, the term signal is often used to indicate risk hypotheses on the use of drugs with data and arguments that support this hypothesis [1]. For our purpose, we refer to the term signal as evidence of an adverse reaction when a given medical product, in many cases a drug, is administered to individuals. Pharmacovigilance authorities aim to retrieve this evidence from the information contained in the databases of spontaneous reports. Pharmacovigilance authorities have the goal of retrieving such evidence by the information contained in their databases. Several issues are intrinsic in the analysis of adrs [5]. There are limitations regarding the quality of the reports that are analysed. Reports are spontaneous thus, in many cases, they are produced by drug users which have not adequate training for deciding whenever the relation between the drug and the symptoms is plausible. There are also limitations linked to the underreporting issue when adrs are not reported by users but they are in act. The percentage of unreported adrs varies from drug to drug and from one adverse reaction to another. Moreover, reports tend to have a non-uniform temporal trend. A drug is more reported after a certain delay from its release to the market, and the attention paid to it changes in time. Moreover, the time to onset, namely the time it takes for a drug's effect to come to prominence, is specific to each drug and depends on

[1] https://www.who-umc.org/research-scientific-development/signal-detection/what-is-a-signal/

additional factors (such as comorbidities or simultaneous administration with other drugs). Each drug must have a specific number of reports, spread in a given duration of time, for determining a plausible adverse reaction of it.

Computational analyses are used to recognize candidate signals that need to be manually investigated to assess their real significance. The majority of them are statistical methods combined with data mining techniques. Statistical disproportionality measures mainly take into account the proportion of associating a drug to a specific adr w.r.t. the complete set of associations related to that drug. Reports are only sent if an individual experiences an adverse reaction possibly caused by a drug administration. The database does not give direct information about the numbers of individuals for which the drug do not cause adverse reactions, neither the number of persons experiencing the reaction but who are not taken the drug. In this context, disproportionality statistics are defined as the ratio between the number of reports in the database and an expected reporting value. Disproportionality measures commonly used for detecting signals of non-causality are the Proportional Reporting Ratio (PRR) [6, 7], the Reporting Odds Ration (ROR) [8], the Information Component (IC) [9], and the Empirical Bayes Geometric Mean (EBGM) [10]. Disproportionality statistics differ in time and resources needed for the computation, the amount of effort and expertise needed to modify them (when required), and the facility with which they can be interpreted by non-experts.

Alternatively, Bayesian methods are also applied. They are the gamma Poisson shrinker [11] and the Bayesian confidence propagation neural networks [9] which is based on a measure called information content (IC)[12]. Moreover, another approach for increasing the power consists in embedding specific characteristics of the drugs and the adverse reaction, (such as therapeutic classification and time since release in the market for drugs, and seriousness and time onset of the adverse reaction) in complex data mining systems [13]. However, these systems highly depend on the type of considered characteristics, and because of their complexity, they can only be applied to small databases or after an

initial selection of candidate signals. Methods can take into account the overall information regarding drugs, adverse reactions present in the entire database. In this perspective, methods can apply subsampling of the input data in order to deal with the sparsity and non-uniformity of the data[14, 15]. Alternatively, a method can divide the overall temporal period of a database into smaller temporal intervals. For example, since reporting rates of adverse events following flu vaccines may vary substantially across years, authors in [3] decided to analyze each year separately. Alternatively, monthly or quarterly intervals can be taken into account [16]. However, the length of the intervals is at the discretion of the analyst who applies his/her expertise depending on a specific drug and adr.

In any case, results may change significantly, independently from the applied statistics, if different temporal lengths are scanned. Several studies have compared the existing computational methodologies [17, 18, 19]. They have performed investigations by comparing the number of obtained false positives, the ratio of false discoveries, the promptness to discover a signal after that the reports regarding it are collects. However, there is no evidence that a specific method outperforms the other approaches in every situation [18, 19]. The *Eudra Vigilance Expert Working Group* of the European Medicines Agency has identified a trade-off between two conflicting goals: methods that generate too many false-positive signals increases the workload of the verification step, however, they also increase the number of found true signals [20].

In this study, we focus the attention on the type of methodologies that make use of intervals by proposing a signal detection method based on variable-length temporal splitting. The main goal is to detect, for each specific drug-adr pair, a set of intervals having different lengths that are representative of the pair under consideration. A set of overlapping intervals are extracted for each drug-adr pair by applying a temporal data-mining approach. The notion of homogeneous interval is introduced. The co-variance coefficient is engaged for detecting cutting points between the intervals in order to extract only homogeneous intervals. Then, a graph theory-based algorithm is applied for retrieving a final set of non-

overlapping intervals. Finally, TEDAR uses the PRR statistics for evaluating the significance of the retrieved intervals. However, any other statistical method can be used for such a purpose.

In this way, the implicit difference between drugs due to their different time on the onset, as well as different timings of the adverse reactions, is mitigated. We compared TEDAR with standard fixed-length interval approaches on the Italian pharmacovigilance system RNF (Rete Nazionale Farmacoviglanza) dataset [2]. We used the ADReCS and PROTECT [20, 21] datasets contained verified drug-adr relations for assessing the performances of TEDAR.

In summary, we proposed a methodology that, for the first time in literature, extends data-mining based methods for the detection of signals by innovatively extracting and scanning intervals of variable length. As a result, the proposed approach detects a greater number of true signals, w.r.t. fixed-length approaches, without significantly increasing the number of false positives. TEDAR also detects adverse reactions more promptly, months before the other methodologies. Moreover, the approach is not dependent on a user-specified parameter, that is the length of the intervals.

TEDAR is freely available at *https://github.com/InfOmics/TEDAR*.

2. The TEDAR methodology

This Section reports a formal description of the proposed method together with the basic notions regarding the pharmacovigilance domain and the data used to validate the proposed method, TEDAR. Table 1 summarizes the main terminology used along the Section.

Given a set of drugs D and a set of adverse reactions A, we define a *drug-adr pair*, on simply a *pair*, a tuple $(drug, adr)$ where $drug \in D$ and $adr \in A$. DA is the combination of all possible $drug - adr$ pairs obtained by associating any drug in D to any adr in A. We define $\mathcal{T}$ as a set of time points within a temporal interval of interest. *Time points* are temporal dates arranged in

D	a set of drugs
A	a set of Adverse Drug Reaction (ADR)
drug-adr pair (simply pair)	an association between a drug and an ADR
((drug, adr),t)	a temporal report of a drug-adr pair
signal	a drug-adr pair supported by a test of riskiness
DA	the combination of all possible drug-adr pairs
$\mathcal{T}$	a set of time points
$R = DA \times \mathcal{T}$	the set of all possible reports in DA
$S : R \to \mathbb{N}$	a survelliance dataset
$\mathcal{R} : R \to \mathbb{N}$	the reporting value associated to a pair
$n_{(drug,adr)}$	number of reports of a given drug-adr pair

Table 1: A summary of the terminology and notation used in this article.

a hierarchical way in order to represent the natural ordering of days, months and years. A *report* is a tuple $r = ((drug, adr), t)$ that represents the detection of the association $(drug, adr) \in DA$ at time $t \in \mathcal{T}$. For instance, the report `((Paracetamol ,Migraine),(2012,12))` means that the association between the drug `Paracetamol` and the adverse reaction `Migraine` was reported in December 2012. The set of all possible reports is referred to as $R = DA \times \mathcal{T}$.

Reports are collected in databases called surveillance. For our purposes, a *surveillance database* S is a function $S : R \to \mathbb{N}$. The database reports the number of distinct individuals for which the pair $(drug, adr)$ is reported at time t. Our method extends the analysis of pairs from single time points to periods of time of fixed length, which we define below through the concept of intervals of time points. Time points can be arbitrary defined as minutes, days, months or years. Considering the frequency of pharmacovigilance reports, we preferred to use months.

Definition 2.1 (Interval of time points). Given a set of time points $\mathcal{T}$, a time points interval $[t, t'] \in \mathcal{T} \times \mathcal{T}$ is specified by the initial and the final time points, t and t' respectively, such that $t \leq t'$, namely t is temporally lower or equal that t' according to a hierarchical ordering of the time points. The length of the interval is defined as $|[t, t']| = |\{\bar{t} : t \leq \bar{t} \leq t'\}|$.

The notion of report can be extended to temporal intervals, such that given

a temporal interval $[t, t']$, $((drug, adr), [t, t'])$ represents the reporting of the $(drug, adr)$ pair in the given interval.

Surveillance databases collect the associations of drugs and adrs that occur for decades. However, each pair appears in a defined period of time called a *timespam* of a drug-adr pair.

Definition 2.2 (Timespan of a drug-adr pair). The *timespan* of a drug-adr pair is an interval $[t_f, t_l]$ such that t_f and t_l are respectively the earliest and the latest time points in which that specific pair is reported in the database.

Definition 2.3 (Reporting value of an association). The reporting value of an association informs about the number of reports for which a given drug is associated with a given adr. Reporting values can refer to single time points or to temporal intervals. A reporting value, for a single time point, is formally defined as a function $\mathcal{R} : R \to \mathbb{N}$, such that $\mathcal{R}((drug, adr), t) = |S((drug, adr), t)|$. Similarly, a reporting value of an association over a time interval $[t, t']$ is defined as $\mathcal{R}((drug, adr), [t, t']) = \sum_{\hat{t} \sqsubseteq [t,t']} S((drug, adr), \hat{t})$, where the operator $\sqsubseteq$ implies that t' is temporally contained in t in the hierarchical order.

For instance, the reporting value in a time point `December 2012` is $\mathcal{R}$`((Paracetamol,Migraine),(2012,12))=10` meaning that in the surveillance database were registered `10` reports associating `Paracetamol` with `Migraine` in `December 2012`. Instead, the reporting value defined on a time interval $\mathcal{R}$`((Paracetamol,Migraine),[(2012,1),(2012,12)])=20` means that 20 reports associate `Paracetamol` with `Migraine` in the entire interval ranging from `January 2012` to `December 2012`.

Definition 2.4 (Proportional Reporting Ratio). Given a drug *drug* and an adverse reaction *adr* in a database S, the Proportional Reporting Ratio (PRR) is computed as

$$PRR(drug, adr) = \frac{x/(x+c)}{b/(b+y)} \tag{1}$$

where x is the number of reports in S for which $drug$ and adr are reported together; b is number of reports for which adr is reported without $drug$; $x+c$ is number of reports for which $drug$ is reported, independently from other factors; and $b+y$ is the total number of reports in which $drug$ is not reported.

The possible values of the $PRR(drug, adr)$ function range from $-\infty$ to $+\infty$.

An high PRR value means that the $(drug, adr)$ pair appears frequently compared to the frequency of $drug$ in the database. A low PRR value means that the $drug$ appears more often associated with other ADRs with respect to adr.

The PRR is conventionally set to ∞ when $b = 0$, i.e. adr is only reported for the $drug$, namely, adr is "specific" to the $drug$.

We have chosen the PRR as a disproportionality measure because it shows good performances in terms of execution time and quality of the results[22]. Besides disproportionality measures, it is necessary to individuate conditions (this is achieved by setting thresholds) that must be satisfied in order to consider the signal as a risky association that needs further countercheck. Thresholds for disproportionality measures are mainly based on the disproportionality statistic itself and on the number of received reports. The choice of thresholds strongly influences the obtained results, which is the reason why we compare different types of thresholds. We have taken into account three different types of commonly used thresholds[17]. One threshold is based on the number of considered cases N that can be greater than or equal to 5. The other two thresholds are based on statistics of the PRR values. Such statistics are the chi-square statistic χ^2 computed for the PRR and the confidence interval. The notion of confidence interval for PRR has been introduced in [6]. The confidence interval is an interval estimation that proposes a range of plausible values for an unknown parameter (i.e. the PRR). The interval has an associated confidence level that the true parameter is in the proposed range [23]. We compute the chi-square statistics as described in the European Medicines Agency (EMA)

guideline [22]. The EMA guidelines suggest a threshold based on the number of cases N greater than or equal to 3. However, the value 3 is only used for active substances contained in medicinal products listed in the additional EMA monitoring list. Products are included in the list only if a post-authorisation safety study (PASS) is requested, i.e. medicinal products authorised in the EU containing a new active substance. Since TEDAR is a generic methodology and none of these medical products is present in the Italian surveillance RNF dataset, we use as said before 5 as value for the threshold for N and for the other thresholds the values reported in their original definitions[6, 23, 22].

Definition 2.5 (Thresholds to define signals). Given an interval $[t, t']$ of a $(drug, adr)$ pair and minimum number of reports within the interval N, the following criteria are applied to detect a signal [16]:

- The lower bound of the 95% PRR confidence interval $\bar{C}$ must be ≥ 1 and $N \geq 5$
- $\chi^2 \geq 4$, $PRR(drug, adr) \geq 2$ and $N \geq 5$

State of the art methods compute PRR statistics by considering a whole timespan as a single interval[3]. In this case, $\mathcal{R}$ collapses all the reports from different time points into a single value without considering temporal trends (i.e. distribution of reports along timespans) in the reporting rates. In order to overcome this limitation, a straightforward method divides timespans in intervals according to a predefined time unit. This procedure was firstly applied in [3], where authors computed signals detection per year. Different drugs come with different amounts of reports. In addition, such reports can be non-homogeneous spread along with the overall timespan. Thus, fixed intervals may be formed by a discontinuous amount of reports which lead to a non-informative leveraging of their information content. Our method is based on the intuition that highlighting only significant variations in the number of reports within a timespan can increase the detection power. In this perspective, it can result useful to

merge time points into intervals, and intervals into larger intervals. State of the art methods perform such operations by producing intervals of fixed length. In contrast to this widely-used way of proceeding, our method for the first time in this domain of application introduces the use of intervals of variable lengths (described in Section 2.1) minimizing the possible overlapping of such intervals (described in Section 2.2) to define the final list of intervals where the measures to detect signals will be applied.

The following steps summarize the main steps that compose the TEDAR methodology:

- Re-scale reports to a selected temporal grain (months);
- Extend time points to overlapping intervals by ensuring homogeneity;
- Retrieve non-overlapping intervals;
- Apply PRR statistics (χ^2 or confidence interval);
- Optionally cross extract signals with existing data of known adverse reactions.

2.1. Homogeneous time points interval

Consecutive time points having similar values of $\mathcal{R}$ are merged to form *homogeneous intervals*. Homogeneous intervals are computed for each drug-adr pair in the database.

We define the homogeneousity of an interval by combination of reporting rate values as their *mean*, *standard deviation* and *coefficient of variation*.

Definition 2.6 (Mean of reporting rate). The mean $\widehat{\mu}((drug, adr), [t, t'])$ is the mean of the values of the reporting rate within a given interval $[t, t']$ for the specific pair $(drug, adr)$. It is defined as:

$$\widehat{\mu}((drug, adr), [t, t']) = \frac{\sum\limits_{t \leq \bar{t} \leq t'} \mathcal{R}((drug, adr), \bar{t})}{|[t, t']|} \tag{2}$$

Definition 2.7 (Standard deviation of reporting rate). The standard deviation $\widehat{\sigma} : (DA \times \mathcal{T} \times \mathcal{T}) \to \mathbb{R}$ is

$$\widehat{\sigma}((drug, adr), [t, t']) = \sqrt{\frac{\sum\limits_{t \leq \bar{t} \leq t'} \left(\mathcal{R}((drug, adr), \bar{t}) - \widehat{\mu}((drug, adr), [t, t'])\right)^2}{|[t, t']|}} \tag{3}$$

where $\widehat{\sigma}((drug, adr), [t, t'])$ is the standard deviation of the association $(drug, adr)$ within the interval $[t, t']$. For each time point $\bar{t}$ of the interval, the mean reporting rate $\widehat{\mu}$ is subtracted to the reporting rate of the drug-adr pair for the time point $\mathcal{R}((drug, adr), \bar{t})$.

Definition 2.8 (Coefficient of variation of reporting rate). Given an interval $[t, t']$, the coefficient of variation $C_V : DA \times \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ is computed as

$$C_V((drug, adr), ([t, t']) = \frac{\widehat{\sigma}((drug, adr), ([t, t']))}{\widehat{\mu}((drug, adr), ([t, t']))} \tag{4}$$

The aim is to obtain a set of *homogeneous* intervals which completely cover a given timespan. To do so, we compare the C_V of each subinterval with the C_V of the timespan multiplied by a tolerance parameter ϵ, $0 \leq \epsilon \leq 1$. ϵ is dynamically defined as $N/n_{(drug,adr)}$, where N the minimal number of cases defined on the threshold of Definition 2.5, and $n_{(drug,adr)}$ is the number of reports n within the timespan of $(drug, adr)$.

Moreover, ϵ allows modulating the number of homogeneous intervals that compose a timespan. For associations with few reports, it is convenient to allow a small number of homogeneous intervals (i.e., high ϵ), while for associations with many reports it is better to allow a greater number of homogeneous intervals (i.e., low ϵ). This enables to better capture the behaviour of the trend of an association.

Definition 2.9 (Homogeneous interval). We say that $[t, t']$ is *homogeneous* , w.r.t a pair $(drug, adr)$, if and only if

$$C_v([t, t'], drug, adr) \leq \epsilon \cdot C_v([t_f, t_l], drug, adr) \tag{5}$$

where $[t_f, t_l]$ is the timespan of the pair $(drug, adr)$.

The coefficient of variation of an interval is compared with a fraction of the coefficient of variation of the timespan. The aim is to produce intervals for which the included time points have a similar reporting value. This notion of homogeneity rules out drops and peaks of reporting values according to the order of the time points.

2.2. Retrieving a minimal set of non-overlapping homogeneous intervals

The previously described procedure produces a set of homogeneous intervals that can overlap within a given timespan. Moreover, some time points within the timespan may not have been included in any of the discovered homogeneous intervals. Given an input surveillance database, such as the RNF database, the proposed approach (TEDAR) scans, for each drug-adr pair reported in the database, all the possible intervals that are candidates to be homogeneous. A refinement procedure, described below, is applied for retrieving a minimal set of homogeneous intervals. Then, the thresholds to the PRR value (see Definition 2.5) are applied. The refinement is applied in order to represent a timespan as a set of homogeneous intervals which do not overlap each other but which cover the entire timespan. The goal is to obtain the set with the minimum number of chosen intervals.

We reduce this problem to find a shortest path in a directed acyclic graph (DAG) $G = (V, E)$. A DAG for each drug-adr association is built. Nodes of the graph represent time points. Initially, every time point contained in the timespan of the drug-adr pair is added to the graph.

Let $succ : \mathcal{T} \mapsto \mathcal{T}$ to be a *successor* function which, given a time point, returns the successive time point in the temporal order. For example, if time points are months, the function returns the month that succeeds the input one. Let $f : \mathcal{T} \mapsto V$ be a function that maps time points to the nodes of the DAG. Given a timespan $[t_f, t_l]$, the number of nodes in G is the cardinality of the

interval $[t_f, t_l]$ plus 1. Thus, the nodes of G are those representing the time points from $f(t_f)$ to $f(succ(t_l))$.

Given two nodes u and v, an edge is represented as an ordered pair (u, v), such that the edge starts from the node u and ends to the node v. An initial set of edges is embedded in the graph, that is $\{(u, v) : \forall t \in [t_f, t_l] \Rightarrow f(t) = u, f(succ(t)) = v\}$.

In this way, we obtain a single path that includes all the time points linked two-by-two by edges. G has also an initial and a final node that are fixed and that are defined by the temporal limits of the timespan in the dataset. Given the starting and ending time points of a homogeneous interval, an edge is added to the DAG by connecting the node that corresponds to the starting point of the interval with the node that corresponds to the time point succeeding the end of the interval. Namely, each input interval $[t, t']$ is encoded by the edge $(f(t), f(succ(t'))$.

A path is an ordered set of nodes $(u_1, u_2, \ldots, u_n)$ from the start node of the DAG, that is $u_1 = f(t_f)$, to the final node of it, that is $u_n = f(succ(t_l))$, such that $(u_i, u_{i+1}) \in E$. The length of a path is given by the edges that are embedded by it, that is the number of nodes of the path minus 1, namely $n-1$. Among all the possible paths of the DAG, a shortest path is a path with the smallest length. Given a shortest path $(u_1, u_2, \ldots, u_n)$, the final set of intervals is obtained by translating edges into intervals. Thus, the final set of interval is given by $\{[t, t'] : f(t) = u_i, succ(f(t')) = u_{i+1}\}$.

We arbitrary chose one of the possible shortest paths, and we extract the set of homogeneous intervals from it. If a given time point is not included in any homogeneous interval, then every shortest path passes on it. By definition, a single time point is a homogeneous interval, thus all the retrieved intervals are homogeneous.

Figure 1 shows an example of the timespan partition process and the choice of the final homogeneous intervals. The timespan ranges from January to Decem-

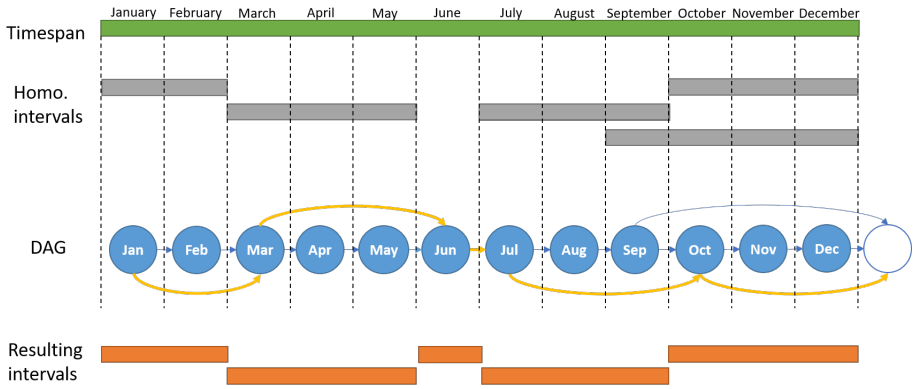

Figure 1: Generation of a DAG of intervals for extracting non-overlapping homogeneous intervals within the timespan of a specific drug-adr pair. Starting homogeneous intervals are displayed as grey rectangles. The initial structure of the DAG is the set of ordered time points (in this case months), represented as nodes (blue nodes) and linked by single edges (blue edges). Initial intervals are embedded in the DAG by adding extra edges from the starting time point to the DAG node consecutive to the one representing the end of the interval (blu edges). For this reason, an extra node is queued to the DAG (white node) in order to represent intervals in which the endpoint is the end of the timespan. Final intervals (orange rectangles) are extracted from one of the possible shortest paths (yellow path) from the start to the end of the DAG.

ber. Initial intervals are computed by the procedure described in the previous Section. Two initial intervals overlap each other (the one from July to September and the one from September to December). The month June is not covered by any interval. A DAG is built in order to contain all the time points (in this case months) from the start to end of the timespan (blue nodes). An additional node (white node) is added to the DAG in order to allow the representation of time intervals which end point is the end of the timespan. The yellow shortest path is chosen for extracting non-overlapping homogeneous intervals in order to completely cover the timespan adding an extra interval representing the month June. Final intervals (orange rectangles) are extracted from one of the possible shortest path (yellow path) from the start to the end of the DAG.

The complete procedure is described in Algorithm 1. The procedure starts

by constructing the DAG G. Then, shortest paths over G are computed, and one of them is arbitrary selected for extracting non-overlapping homogeneous intervals.

Algorithm 1: MinimalHomogeneousIntervalSet

Data: $\mathcal{R}$, a drug-adr pair $(drug, adr) \in DA$, and $\epsilon \in \mathbb{R}$ with $0 \leq \epsilon \leq 1$.

Result: A minimal set $\mathcal{H} = \{[t_0^s, t_0^e], \ldots [t_n^s, t_n^e]\}$ of *homogeneous intervals* such that $t_0^s = \min_{\mathcal{R}((drug,adr),t)>0} t$, $t_n^e = \max_{\mathcal{R}((drug,adr),t)>0} t$, and for each $0 \leq i < n$ we have $succ(t_i^e) = t_{i+1}^s$.

begin

- $t_f \leftarrow \min_{\mathcal{R}((drug,adr),t)>0}$
- $t_l \leftarrow \max_{\mathcal{R}((drug,adr),t)>0}$
- $V \leftarrow \{v_t : t_f \leq t \leq succ(t_l)\}$
- $E \leftarrow \emptyset$
- **for** $v_t \in V \setminus \{v_{succ(t_l)}\}$ **do**
  - $V_{\geq v_t} \leftarrow \{v_{t'} : t \leq t' \leq t_l\}$
  - **for** $v_{t'} \in V_{\geq v_t}$ **do**
    - **if** $C_v([t, t'], drug, adr) \leq \epsilon \cdot C_v([t_f, t_l], drug, adr)$ **then**
      - $E \leftarrow E \cup \{(v_t, v_{succ(t')})\}$
- **let** $v_{t_0} \ldots v_{t_{n+1}}$ be a shortest path from v_{t_f} to $v_{succ(t_l)}$ in $G = (V, E)$ with $t_0 = t_l$ and $t_{n+1} = succ(t_l)$.
- $\mathcal{H} \leftarrow \{[t_i, t_i^e] : 0 \leq i \leq n \text{ and } succ(t_i^e) = t_{i+1}\}$
- **return** $\mathcal{H}$

Level	Term
System Organ Class (SOC)	Gastrointestinal disorders
High Level Group Term (HLGT)	Gastrointestinal signs and symptoms
High Level Term (HLT)	Nausea and Vomiting symptoms
Preferred Term (PT)	Nausea
Lowest Level Term (LLT)	Feeling queasy

Table 2: An example of the five levels of a MedDRA term.

3. Applying the TEDAR model to a real adverse drug reaction surveillance databases

3.1. Data sources

We use as case study the surveillance database, named RNF (Rete Nazionale Farmacovigilanza), released by the Italian authority AIFA (Agenzia Italiana del FArmaco) (http://www.agenziafarmaco.gov.it/). The RNF database contains reports of adrs issued by all the Italian regions. In more details, it has $170K$ reports, $280K$ adrs, and $375K$ treatments. Despite the fact the first report dates back to the year 1985, most of the reports have been produced after the year 2007. Moreover, the information contained in the reports collected from the year 2008 to now, in general, are more reliable due to a re-organization of the national pharmacovigilance centres. Thus, we decided to take into account reports produced since January 2008. Adrs are encoded according to the MedDRA (Medical Dictionary for Regulatory Activities) terminology [24], which consists of a large set of terms structured into five hierarchical levels, shown with an example in Table 2.

System Organ Classe (SOC) is the highest level of adr terminology, terms here are distinguished by anatomical or physiological systems, aetiology or purpose. The hierarchy is multiaxial, i.e. for example, a PT (Preferred Term) term can be grouped in one or more HLT (High Level Term) terms, but it belongs to only one primary SOC (System Organ Class). In our analysis we used the System Organ Classes (SOCs) instead of lower-level terms [3].

The data in RNF meets the requirements of the ICH M5 standard [25].

Such standard solves the issue of lack of internationally harmonised standards related to core sets of medicinal product information and medicinal product terminology. It contains three main entities: (i) the *medicinal products* that are the packages on sale for a period defined by the authorization date and by the possible date of withdrawal from the market (e.g. `ASPIRIN®*20 TABLETS 500 mg`); (ii) *ATC classes* that divides medicinal products into groups based on their chemical and therapeutic characteristics, and of the body part on which they act, managed by the WHO Collaborating Centre for Drug Statistic Methodology; and (iii) the *pharmaceutical products* containing combinations of active ingredients that compose the medicinal (e.g. `acetylsalicylic acid/paracetamol` are the active ingredients that compose `Doloflex`). In our analysis, we refer to pharmaceutical products and we make no distinction between pharmaceutical products with the same combinations of active ingredients. For the sake of simplicity, we refer to them with the term *drug*.

Signals are collected in datasets [2, 3, 4]. and constantly updated. Among them, we used for validation two well established European and USA datasets: PROTECT [20] and ADReCS [21]. PROTECT has been developed by the European Medicines Agency[26]. The database contains 64,498 confirmed drug-adr associations. It also reports information for each drug-adr pair such as the MedDRA codes, the age group, the gender, the clinical trial, the post-marketing surveillance and the causality (that providing information on the strength of causal association). We extracted only the established (causality equal to 0) or very probable (causality equal to 1) associations while those unclear or unassailable (causality equal to 2 or 3) were removed. The number of records remaining after this filtering phase is 63.826.

ADReCS is managed by the U.S. National Library of Medicine (NLM). The drug-adr information of ADReCS is mainly extracted from the drug labels in the DailyMed (http://dailymed.nlm.nih.gov/dailymed/about.cfm), SIDER2[27] and the U.S. Food and Drug Administration (US FDA)[28]. ADReCS contains 27,328 drug-adr associations, with 1,697 distinct drugs.

The RNF dataset contains 3,043 drugs and 27 ADRs, which are put in relations by 37,393 pairs. To adapt the data from those validation datasets to our experiments several manipulation steps were carried out for both drugs and adrs. The RNF dataset uses Italian terms, while PROTECT and ADReCS use the English language. We mapped drugs in RNF to those in PROTECT and ADReCS by using the active ingredient of the drugs, rather than their commercial name, and by applying text mining based on the Levenshtein's strings distance. Drugs composed of the exact combination of multiple active ingredients are seen as a unique drug. On the contrary, if two drugs differ for at least one ingredient, they are considered separately. For what concerns adverse reactions, we mapped them via their SOC codes. The switch from PTs to SOC codes, in PROTECT and ADReCS, produces a collapse of all the drug-adr pairs relative to different PTs that correspond to a single SOC code. By grouping by drugs and SOCs in PROTECT we obtained 7.544 associations and 11.685 in ADReCS. By mapping the PROTECT database to the RNF dataset, we extracted 505 drugs, 27 ADRs, and 5,460 pairs. Instead, by mapping the ADReCS database to the RNF dataset, we obtained 519 drugs, 27 ADRs, and 7,678 pairs.

3.2. Experimental results

We compare TEDAR with the standard signal detection methods which exploit disproportionality by using the Proportional Reporting Ratio (PRR)[3]. The standard methods take into account intervals having yearly, monthly or quarterly fixed length. Based on the length of the intervals, we refer to them as *PRR yearly*, *PRR quarterly*, *PRR monthly*. The three approaches assign a disproportionality score to each drug-adr pair present in the input dataset. Then, they apply a threshold based on the Chi-squared statistic or on the confidence interval in order to obtain the final set of pairs that are considered signals (see Section 2 and Definition 2.5).

We applied the standard signal detection methods and TEDAR to the RNF database considering all associations $(drug, adr)$ in a timespan of 10 years

[(2008,1), (2017,12)] as explained in Section 3.1. We used two measures for evaluating the performance of the compared methods: the F1-score[17] and the early signal detection power[26].

We computed the F1-score as a combination of precision and recall based on the presence of the retrieved signals in the reference datasets composed by the union of ADReCS and PROTECT. From the two reference standard datasets, we selected only the drug-ard pairs for which a minimum number of reports equal to 5 is reported in RNF. The excluded pairs did not have enough support in the RNF dataset to be detected as signals. In this way, we obtained 5,565 and 3,952 pairs for ADReCS and PROTECT, respectively. By merging these two datasets, we obtained a set of 8,568 drug-adr pairs of which 1,040 are in common between ADReCS and PROTECT. From now to below, the merged set is referred to as the reference dataset.

F1-score is a comprehensive evaluation metric that represents a harmonic mean between sensitivity and precision. It is evaluated as $\frac{TP}{TP+\frac{1}{2}(FP+FN)}$, where TP (true positive) is the set of drug-adr pairs that are detected by a given approach and that are known to be real signals, namely they are reported in the reference; FP (false positive) is the set of pairs that are detected as signals but they are not reported in the reference; FN (false negative) is the set of pairs that are not detected by the given approaches but are present in the reference.

The calculation of the F1-score mainly depends on the number of pairs that are detected by a method and on their percentage that is reported in the reference. Table 3 reports the number of detected signals for all the compared methods by using the confidence interval and the Chi-squared statistics. Fixed-length approaches and the default TEDAR solution discard intervals supported by less than 5 reports, which means that the parameter N is set to be 5. A further experiment, obtained by setting $N = 10$ for TEDAR, is also reported. For this experiment, we also discarded from the reference all the drug-adr pairs having less that 10 reports, leading to a total of 6,860 items.

TEDAR and *PRR quartely* are the approaches that extract the higher num-

	Chi-squared		Confidence interval	
	Total	in reference (TP) (8,568 ref.)	Total	in reference (TP) (8,568 ref.)
PRR yearly	6,087	2,786	8,119	3,817
PRR quarterly	6,572	3,279	7,784	3,942
PRR monthly	5,890	3,123	6,532	3,461
PRR TEDAR	**7,308**	**3,538**	**8,732**	**4,220**
PRR TEDAR ($N = 10$)	3,204	1,783 (6,860 ref.)	4,005	2,223 (6,860 ref.)

Table 3: Number of total signals found by each approach by using the confidence interval and the Chi-squared statistic. Best performances are written in bold.

ber of signals. This is a required feature by the European Medicines Agency for characterizing the goodness of a detection method since false positive signals increase the workload of the verification step, however, they also increase the number of found true signals[20]. TEDAR and *PRR quartely* are also the approaches that detect the higher number of pairs that are also reported in the reference. This fact supports the choice of pharmacovigilance agencies to use, among the standard methods, the *PRR quartely*, as is the case of the Italian Pharmacovigilance Agency. By using the confidence interval, the compared approaches are more suitable in finding pairs that are also reported in the reference dataset. In fact, they detect an average of 45% of pairs of the reference, in contrast to an average of 37% if the Chi-squared statistic is used. Finally, Table 3 shows that TEDAR compared to *PRR quartely* and other standard methods retrieve more signals present in the reference.

The use of $N = 10$ naturally reduces the number of pairs to be evaluated. In fact, the number of signals found by TEDAR, with such a parameter, is relative low. Moreover, all the 3,204 signals detected by using the Chi-squared statistics, as well as all the 2,223 signals detected via the confidence intervals, are enclosed in the items retrieved by TEDAR with $N = 5$. This aspect reflects the common practice of setting $N = 5$ in pharmacovigilance approaches.

All the compared methods have a relatively high number of signals that are not reported in the reference dataset. The pairs that are not detected by

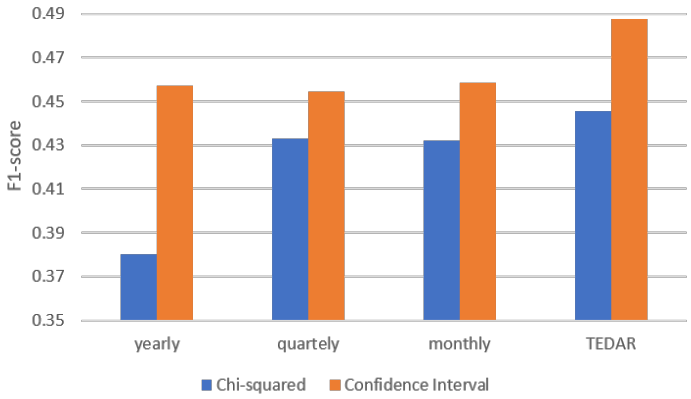

Figure 2: F1-score of the four compared approaches by using the Chi-squared and confidence interval statistics.

any method tend to have fewer reports. Their reports are more distant and, therefore, more spread in the timespan. All missing pairs could be false positives or true negatives. They can potentially constitute true signals not reported in the reference dataset. Unfortunately, the currently available reference datasets do not report drug-adr pairs for which the manual validation have discarded them as true signals. Namely, the reference dataset does not report true negative associations. Considering all these observations, we chose the F1-score as a performance measure, since its calculation is not based on the number of true negatives. Furthermore, as reported in [16], we do not expect the accuracy to be high, and therefore even small differences F1 scores allow us to evaluate which of the compared methods is more promising.

Figure 2 shows the F1-scores of TEDAR, *PRR yearly*, *PRR quarterly* and *PRR monthly*. TEDAR outperforms all standard approaches confirming that variable-length intervals fit better the data in order to get signals compared to the use of fixed-length interval lengths. Regarding the results with the Chi-squared statistic, the *PPR quarterly* and *PRR monthly* have comparable performances. On the contrary, the three standard approaches have similar performance when the confidence interval is used, and they differ from TEDAR.

Pharmacovigilance analyses are used to detect as early as possible signals. This is a crucial factor in choosing one method over another. It allows for preventive actions that reduce risks for patients, such as removing a drug from the market as soon as possible. Furthermore, this performance evaluation is a more stable and objective measure of comparison since it does not refer to the goodness of the existing reference datasets.

We computed the early signal detection power of the standard methods and TEDAR by taking into account the average time needed for detecting a signal since the first report appeared into the pharmacovigilance datasets reporting the drug-ard pair. Given a drug-adr pair detected by a method, we calculated the number of months that intercluded between the first reporting of the pair in the input dataset and the time that the method was able to detect the signal.

A detection method requires a given amount of reports for detecting a $(drug, adr)$ pair as a potential signal. Moreover, reports have to be organized in a specific period of time. For example, the *PRR monthly* method is not able to detect signals if the reports that support their candidature do not cluster in months. The same applies to quarterly and yearly methods in the different timescale (quarters and years). Moreover, clusters may be formed after an initial period where the pair is reported sporadically. Thus, the fixed-length methods compared to TEDAR are not able to promptly detect such signals because no clusters within the right time intervals are formed.

Figure 3 reports average delays (and the standard deviation) of the compared methods and show that TEDAR, independently by the statistics, notable outperforms any compared methods. TEDAR requires on average 22 months compared to 30 or 35 to detect the signal when the Chi-squared statistic is used. The delays are lower than 12 months and the gap among the compared methods increases (they detect signals starting from the 27 to 35 months) if the confidence interval is used. The confidence interval results in the best statistic to be used for what concerns the early detection of signals.

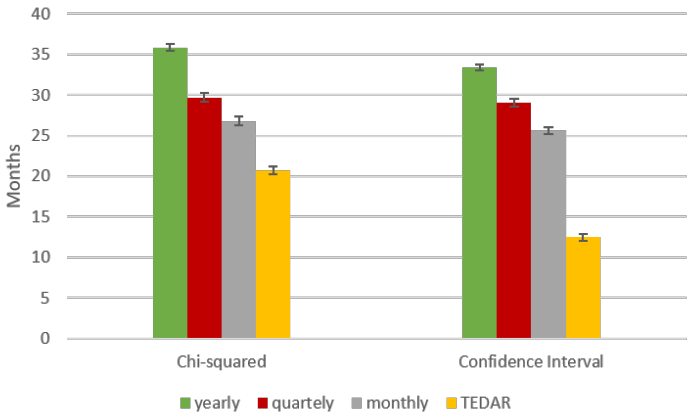

Figure 3: Early detection power of the compared methods computed as the average time (and standard deviation) in months to detect adverse drug reactions considering the different statistics.

Next, we report computational time required to execute all compared algorithms, together with the time required to import RNF data to Redis database. Tests were performed on top of a commodity machine equipped with an Intel(R) Core(TM) i7 CPU (1.80GHz), 8 GBs of RAM and an SSD hard drive. The time required to load the data and compute statistics on the reports (to speed the PRR calculation up) was about 7 hours. Unlike fixed-length interval methodologies, TEDAR required an additional time of approximately 2 hours to compute and extract non-overlapping homogeneous intervals set. On the other hand, the application of the disproportionality measure to time intervals required an almost similar time of about 1 hour for all the tested methodologies (*PRR yearly*, *PRR monthly*, *PRR quarterly* and *TEDAR*), and it was proportional to the number of intervals extracted from the timespan.

Compared to fixed length interval methodologies, the choice of the time points granularity in TEDAR is up on the users. This is the main limitation of the method, although we suggest to define time points as months to obtain consistent performance.

Finally, we investigated the intervals retrieved by TEDAR in order to deeply understand the outperformance of TEDAR and give one more justification of the need to use variable-length intervals in pharmacovigilance signal detection methods. We show the analysis made by using the confidence interval statistic, however, a similar trend is notable for what concerns the χ^2 statistic (Figures not reported here).

Figure 4 shows, for each specific interval length, the number of non-overlapping homogeneous intervals retrieved by TEDAR having such a length. The length is given in months since it is the chosen unitary time point. The chart arranges the retrieved count by putting the interval length on the x-axis and the corresponding number of intervals having such a length on the y-axis. Lengths are sorted in decreasing order from left to right. The total number of retrieved intervals is 37k, of which 11k are one month long. The curve decreases exponentially. An exception is a group of 453 intervals that are 120 months long and that form a peak at the tail of the histogram.

Figure 5 shows the intervals filtered by applying the confidence interval statistics. They are in total 8,732 of which 6,800 are one month long. The curve formed by the histogram follows the exponential decay trend of Figure 4, and the peak of 453 non-filtered intervals of 120 months are lowered to 108 by the applied statistics.

The similar trends of the two histograms highlight that the core technique for the extraction intervals is useful for producing adr candidates is given by the retrieving of the variable length intervals, which is the main novelty of the proposed approach. In fact, the application of PRR to the retrieved intervals does not reshape the histogram, but it only filters out a proportional amount of intervals for each length. Such a behaviour validates the intuition that adr signals are supported by intervals having variable lengths.

More strong evidence of such behaviour is due to the similar trend that emerges when only TP signals are extracted from the histograms. The evaluation of TP is made in accordance with the reference data set described in

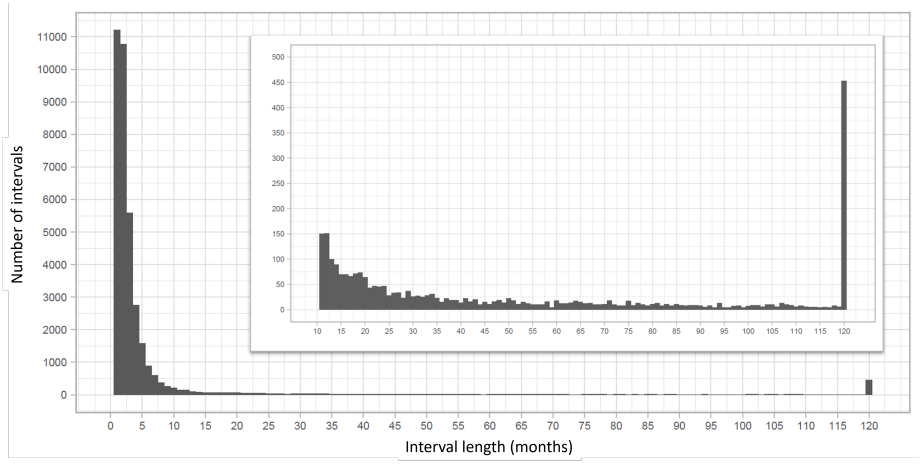

Figure 4: Number of non-overlapping intervals retrieved by TEDAR grouped by the length of the intervals (in months). The inner chart shows a re-scaled zoom of the whole histogram by only showing interval lengths from 10 to 120 months.

Section 3.1. Intervals corresponding to drug-adr pairs reported in the reference data set described are considered True Positive (TP). The trend is reported in Figure 6. Validated signals correspond to intervals of different lengths. The shorter is the interval produced by TEDAR, the less are the chance to detect the signal for a method working with intervals of longer lengths. For example, a signal that requires at least an interval of two months for being detected can hardly be retrieved via a fixed-length window of one month. Such evidence is clear for short lengths of the intervals, however, the histogram values highlight that the positive effects of variable-length intervals are present at any temporal scale.

4. Conclusions

We have proposed a method, TEDAR, for detecting signals of adverse drug reactions from pharmacovigilance systems based on spontaneous reporting. We have compared TEDAR to current methodologies based on a fixed interval length, that split the lifetime of a drug into months, quarters and years. TEDAR

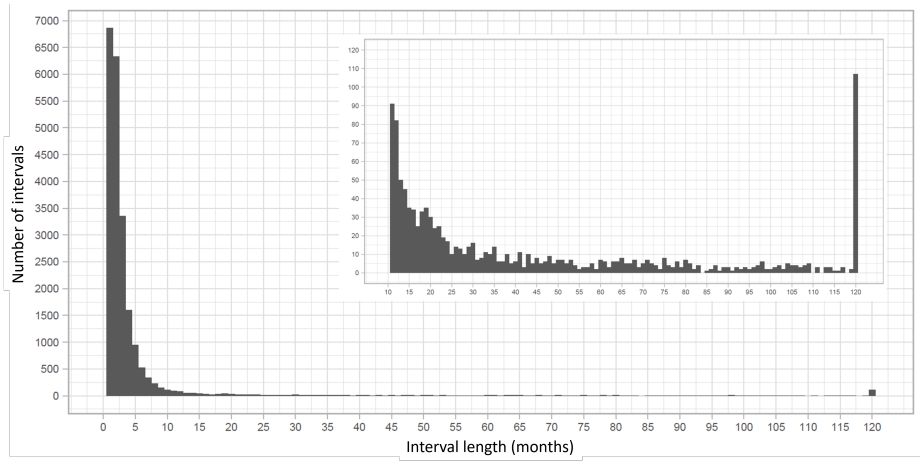

Figure 5: Number of non-overlapping intervals retrieved by TEDAR that passed the confidence interval statistic. The length is calculated in month. The inner chart shows a re-scaled zoom of the whole histogram by only showing interval lengths from 10 to 120 months.

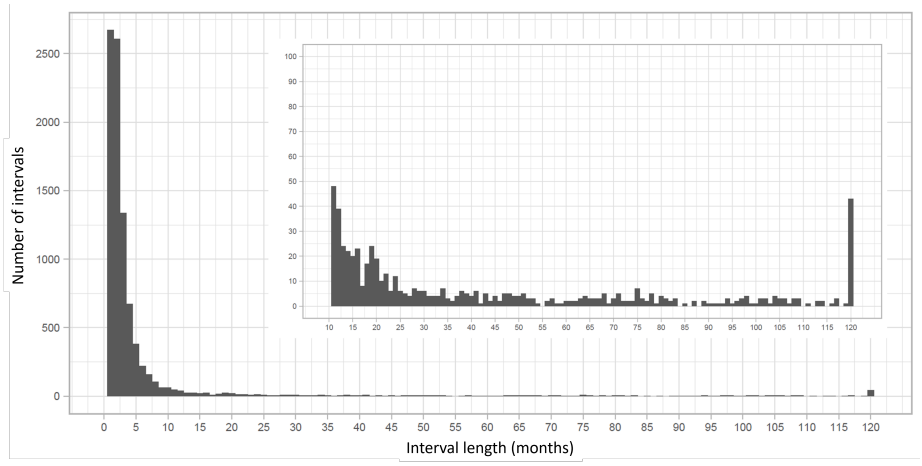

Figure 6: Number of non-overlapping intervals retrieved by TEDAR that passed the confidence interval statistic. The length is calculated in month. The results are obtained by using the reference dataset. The inner chart shows a re-scaled zoom of the whole histogram by only showing interval lengths from 10 to 120 months. The histogram reports only intervals classified as TP in accordance with the used reference data set described in Section 3.1.

resulted in a more flexible approach since it does not depend on user preferences, namely the length of the temporal intervals that are scan for detecting the signals. The results showed that the use of a variable-length approach increases the effectiveness in detecting signals of adverse drug reactions. In fact, by applying different statistics for measuring the strength of the signal, TEDAR outperformed the other approaches by showing always a better F1-score. Moreover, TEDAR was able to detect signals months before the other methods can do. An analysis of the temporal intervals which produce signals showed that their length is not a fixed value but it varies according to exponential distributions. Thus, methods that focus only on a given interval length miss the detecting of a relatively large portion of the whole set of signals.

References

[1] D. R. Poudel, P. Acharya, S. Ghimire, R. Dhital, R. Bharati, Burden of hospitalizations related to adverse drug events in the USA: a retrospective analysis from large inpatient database, Pharmacoepidemiology and drug safety 26 (6) (2017) 635–641.

[2] AIFA Official Website, `http://www.agenziafarmaco.gov.it/content/rete-nazionale-di-farmacovigilanza`.

[3] Y. Cai, J. Du, J. Huang, S. S. Ellenberg, S. Hennessy, C. Tao, Y. Chen, A signal detection method for temporal variation of adverse effect with vaccine adverse event reporting system data, BMC medical informatics and decision making 17 (2) (2017) 76.

[4] W. Zhou, V. Pool, J. K. Iskander, R. English-Bullard, R. Ball, R. P. Wise, P. Haber, R. P. Pless, G. Mootrey, S. S. Ellenberg, et al., Surveillance for safety after immunization: vaccine adverse event reporting system (VAERS)—United States, 1991–2001, MMWR Surveill Summ 52 (1) (2003) 1–24.

[5] G. J. Dal Pan, M. Lindquist, K. Gelperin, Postmarketing spontaneous pharmacovigilance reporting systems, Pharmacoepidemiology (2019) 165–201.

[6] S. Evans, P. C. Waller, S. Davis, Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports, Pharmacoepidemiology and drug safety 10 (6) (2001) 483–486.

[7] Y. Ji, F. Shen, J. Tran, A Multi-relational Association Mining Algorithm for Screening Suspected Adverse Drug Reactions, in: 2014 11th International Conference on Information Technology: New Generations, IEEE, 2014, pp. 407–412.

[8] K. J. Rothman, S. Lanes, S. T. Sacks, The reporting odds ratio and its advantages over the proportional reporting ratio, Pharmacoepidemiology and drug safety 13 (8) (2004) 519–523.

[9] A. Bate, M. Lindquist, I. R. Edwards, S. Olsson, R. Orre, A. Lansner, R. M. De Freitas, A Bayesian neural network method for adverse drug reaction signal generation, European journal of clinical pharmacology 54 (4) (1998) 315–321.

[10] A. Bate, S. Evans, Quantitative signal detection using spontaneous ADR reporting, Pharmacoepidemiology and drug safety 18 (6) (2009) 427–436.

[11] W. DuMouchel, Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system, The American Statistician 53 (3) (1999) 177–190.

[12] F. De Pretis, B. Osimani, New insights in computational methods for pharmacovigilance: E-synthesis, a bayesian framework for causal assessment, International journal of environmental research and public health 16 (12) (2019) 2221.

[13] C. Fukazawa, Y. Hinomura, M. Kaneko, M. Narukawa, Significance of data mining in routine signal detection: analysis based on the safety signals

identified by the FDA, Pharmacoepidemiology and drug safety 27 (12) (2018) 1402–1408.

[14] I. Ahmed, A. Pariente, P. Tubert-Bitter, Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions, Statistical methods in medical research 27 (3) (2018) 785–797.

[15] M. Sardella, C. Lungu, Evaluation of quantitative signal detection in EudraVigilance for orphan drugs: possible risk of false negatives, Therapeutic advances in drug safety 10 (2019) 2042098619882819.

[16] M. Lerch, P. Nowicki, K. Manlik, G. Wirsching, Statistical signal detection as a routine pharmacovigilance practice: effects of periodicity and resignalling criteria on quality and workload, Drug safety 38 (12) (2015) 1219–1231.

[17] G. Candore, K. Juhlin, K. Manlik, B. Thakrar, N. Quarcoo, S. Seabroke, A. Wisniewski, J. Slattery, Comparison of statistical signal detection methods within and across spontaneous reporting databases, Drug safety 38 (6) (2015) 577–587.

[18] G. Park, H. Jung, S.-J. Heo, I. Jung, Comparison of Data Mining Methods for the Signal Detection of Adverse Drug Events with a Hierarchical Structure in Postmarketing Surveillance, Life 10 (8) (2020) 138.

[19] Y. Ding, M. Markatou, R. Ball, An evaluation of statistical approaches to postmarketing surveillance, Statistics in Medicine 39 (7) (2020) 845–874.

[20] A. F. Wisniewski, A. Bate, C. Bousquet, A. Brueckner, G. Candore, K. Juhlin, M. A. Macia-Martinez, K. Manlik, N. Quarcoo, S. Seabroke, et al., Good signal detection practices: evidence from IMI PROTECT, Drug safety 39 (6) (2016) 469–490.

[21] M.-C. Cai, Q. Xu, Y.-J. Pan, W. Pan, N. Ji, Y.-B. Li, H.-J. Jin, K. Liu, Z.-L. Ji, ADReCS: an ontology database for aiding standardization and

hierarchical classification of adverse drug reaction terms, Nucleic acids research 43 (D1) (2014) D907–D913.

[22] EudraVigilance Expert Working Group, Guideline on good pharmacovigilance practices (GVP), Module IX Addendum I. Accessed 4 Aug 2016. https://www.ema.europa.eu/en/documents/scientific-guideline/draft-guideline-good-pharmacovigilance-practices-gvp-module-ix-addendum-i-methodological-aspects_en.pdf.

[23] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaä, L. E. Meester, A Modern Introduction to Probability and Statistics: Understanding why and how, Springer Science & Business Media, 2005.

[24] MedDRA MSSO, About MedDRA, in: http://www.meddramsso.com/public_about_meddra.asp, 2010.

[25] ICH M5 Expert Working Group, Data elements and standard for drug dictionaries M5, in: http://www.ich.org, 2005.

[26] Y. Alvarez, A. Hidalgo, F. Maignen, J. Slattery, Validation of Statistical Signal Detection Procedures in EudraVigilance Post-Authorization Data, Drug Safety 33 (6) (2010) 475–487. doi:10.2165/11534410-000000000-00000.
URL https://doi.org/10.2165/11534410-000000000-00000

[27] M. Kuhn, I. Letunic, L. J. Jensen, P. Bork, The SIDER database of drugs and side effects, Nucleic Acids Research 44 (D1) (2015) D1075–D1079. arXiv:https://academic.oup.com/nar/article-pdf/44/D1/D1075/16661270/gkv1075.pdf, doi:10.1093/nar/gkv1075.
URL https://doi.org/10.1093/nar/gkv1075

[28] U. S. Food, Drug Administration US FDA, Guidance for the In Vitro Portion of Bioequivalence Requirements for Metaproterenol Sulfate and Albuterol Inhalation Aerosol.