FP7-ICT-2007-1

www.euadr-project.org

D3.3 Description of the Data Mining Algorithms and Data Mining Software for Local Signal Generation

WP3 – Signal generation

V1.2

Final Version

Lead beneficiary: EMC

Date: 19-09-2011

Nature: R,P

Dissemination level: CO

TABLE OF CONTENTS

Document History

Name	Date	Version	Description
Miriam Sturkenboom, Preci Coloma	11-01-2011	0.1	
Martijn Schuemie, Preci Coloma, Huub Straatman, Justin Matthews, Mariam Molokhia, David Prieto, Ron Herings, Silvana Romio, Lorenza Scotti, Anni Fourrier, Franz Thiessard, Gianluca Trifero, Johan van der Lei, Miriam Sturkenboom	29-07-2011	1.0	Draft, for internal review
Martijn Schuemie, Scott Boyer, Ernst Ahlberg Helgee, Laura Furlong	09-09-2011	1.1	Final draft after internal review.
Martijn Schuemie, Justin Matthews	19-09-2011	1.2	Final version, after consortium review

Definitions

Partners of the EU-ADR Consortium are referred to herein according to the following codes:

EMC - Erasmus University Medical Center (Netherlands) – Coordinator
FIMIM - Fundació IMIM (Spain) – Beneficiary
UPF - Universitat Pompeu Fabra (Spain) – Beneficiary
UAVR - University of Aveiro – IEETA (Portugal) – Beneficiary
NEUROLESI - IRCCS Centro Neurolesi "Bonino-Pulejo" (Italy) – Beneficiary
UB2 - Université Victor-Segalen Bordeaux II (France) – Beneficiary
LSHTM - London School of Hygiene & Tropical Medicine (UK) – Beneficiary
AUH-AS - Aarhus University Hospital, Århus Sygehus (Denmark) – Beneficiary
AZ - AstraZeneca R&D (Sweden) – Beneficiary
UNOTT - University of Nottingham (UK) – Beneficiary
UNIMIB - Università di Milano-Bicocca (Italy) – Beneficiary
ARS - Agenzia Regionale di Sanità (Italy) – Beneficiary
PHARMO - PHARMO Coöperation UA (Netherlands) – Beneficiary
PEDIANET -Societa' Servizi Telematici SRL (Italy) – Beneficiary
USC - University of Santiago de Compostela (Spain) – Beneficiary
TAU - Tel-Aviv University (Israel) – Subcontractor
SIMG - Health Search - Italian College of General Practitioners (Italy) – Subcontractor
ICL - Imperial College London (UK) – Subcontractor

- **Grant Agreement**: The agreement signed between the beneficiaries and the European Commission for the undertaking of the EU-ADR project (ICT-215847).
- **Project**: The sum of all activities carried out in the framework of the Grant Agreement by the Consortium.
- **Work plan**: Schedule of tasks, deliverables, efforts, dates and responsibilities corresponding to the work to be carried out for the EU-ADR project, as specified in Annex I to the Grant Agreement.
- **Consortium**: The EU-ADR Consortium, conformed by the above-mentioned legal entities.

1. Executive Summary

A wide range of signal detection methods have been evaluated in the EU-ADR project, including methods derived from methods used in spontaneous report databases, methods from epidemiology, and methods specifically designed for longitudinal data. Several novel methods have been developed within the EU-ADR project: the Longitudinal Gamma Poisson Shrinker (LGPS)[1] is an adaptation of a method from spontaneous report systems to longitudinal data, Longitudinal Evaluation Of Profiles of Adverse Reactions to Drugs (LEOPARD)[1] was developed to detect protopathic bias, and the Bayesian Hierarchical Model was developed to utilize the hierarchy in drug taxonomy.

The pre-processing and aggregation needed for all methods was implemented in the software framework of Jerboa, a Java program developed in EU-ADR that is run locally by each database. The output of Jerboa was collected at a central site for further analysis.

In order to evaluate the performance of methods, a reference set was created using ADRs reported in literature as the gold standard: Drug-event combinations that are described and evaluated in many publications are considered positive signals, drug-event combinations with no evidence in literature are deemed negative signals. The set was limited to drug-event pairs for which there was enough exposure to detect a relative risk of 4. In total 44 positive and 50 negative signals were identified.

Method performance was primarily measured using the area under the receiver operator curve. Data from the different databases was combined by either pooling the data, or by use of meta-analysis techniques.

All methods performed better than random baseline, and there was not much difference in the performance of the different methods. LEOPARD filtering in general improved performance. LGPS and case-control in combination with LEOPARD slightly outperformed the other methods. The highest measured area under the curve was 0.83.

2. Introduction

Modern drug legislation was promped over 40 years ago due to serious adverse effects resulting from the treatment with thalidomide [2]. Since then, the mainstay of drug safety surveillance has been the collection of spontaneous Adverse Drug Reactions (ADRs)[3, 4]. The current and future challenges of drug development and drug utilization, and a number of recent high-impact drug safety issues (e.g. rofecoxib (Vioxx) and SSRIs) require re-thinking of the way safety monitoring is conducted[5]. It has become evident that adverse effects of drugs may be detected too late, when millions of persons have already been exposed. The need to change drug safety monitoring is underlined in the current public consultation about the future of pharmacovigilance in the EU.

Pharmacovigilance is the study of the safety of marketed drugs under the practical conditions of clinical usage in large communities. The timely discovery of unknown or unexpected ADRs is one of its major challenges, because most of the drugs enter the market with less than 3000 exposed subjects, implying that reactions occurring with rates lower than 1/1000 could easily remain undetected for long periods of time. Post-marketing Spontaneous Reporting Systems (SRSs) for suspected ADRs have been a cornerstone to detect safety signals in pharmacovigilance[6]. Although many ADRs were detected by SRSs, these systems have inherent limitations that hamper signal detection[7]. The major weakness is that these systems depend entirely on the ability of a physician to, first, recognize an adverse event as being related to the drug. Subsequently, the physician needs to actually report the case to the local spontaneous reporting database. The greatest limitations, therefore, are under-reporting and biases due to selective reporting[8]. Investigations have shown that the percentage of ADRs being reported varies between 1 and 10%[9-11]. These problems may lead to underestimation of the significance of a particular reaction and delay in signal detection, as well as spurious detections[12].

In EU_ADR, an alternative approach towards the detection of ADR signals has been developed with the objective of overcoming the shortcomings of SRSs and providing a solid basis for large-scale monitoring of drug safety. Rather than relying on the physician's capability and willingness to recognize and report suspected ADRs, the system will systematically calculate the occurrence of disease (potentially ADRs) during specific drug use based on data (time-stamped exposure and morbidity data) available in electronic patient records. Europe plays a leading role in the development and use of electronic patient records[13, 14]. As a result, a number of European Electronic Healthcare Record (EHR) databases are available. Appropriate monitoring and use of these databases has an enormous potential for earlier detection of ADR signals[15, 16].

Longitudinal healthcare databases containing medical records and administrative claims have long been used to characterize healthcare utilization patterns, monitor patient outcomes, and carry out formal pharmacoepidemiological studies. With regards to drug safety surveillance, such databases have been most commonly used to confirm or refute potential signals flagged by spontaneous reporting or other surveillance systems. EHR databases are appealing for safety signal evaluation because of their large size, accurate exposure capture and broad population coverage. Since data are routinely collected for

other purposes and, hence, incurs no additional cost, these databases offer the advantage of efficiency in the conduct of drug safety studies. This is in addition to their ability to provide practical clinical data culled from real-world settings. In the last few years several international collaborations have ventured beyond using EHR databases for signal confirmation to developing EHR-based drug safety signal detection systems.

Similar efforts

EU-ADR has similar aims as several initiatives in the USA such as the OMOP and Sentinel that have recently started. The Food and Drug Administration (FDA) in the USA has started the Sentinel Initiative in recognition of the need to use innovative methods to monitor FDA-regulated products and to enhance public health safety by secondary use of anonymised health data. In the autumn of 2007, Congress passed the FDA Amendments Act (FDAAA), mandating the FDA to establish an active surveillance system for monitoring drugs that uses electronic data from healthcare information holders. The Sentinel initiative is the FDA's response to that mandate. Its goal is to build and implement a new active surveillance system that will eventually be used to monitor all FDA-regulated products in a total of at least 100 million patients.(http://www.fda.gov/Safety/FDAsSentinelInitiative/default.htm)

The Observational Medical Outcomes Partnership (OMOP) is a public-private partnership designed to help improve the monitoring of drugs for safety. The partnership is developing and testing research methods that are feasible and useful to analyze existing healthcare databases to identify and evaluate safety and benefit issues of drugs already on the market (http://omop.fnih.org/).

Recently, in Europe the PROTECT project was started (http://www.imi-protect.eu). It is funded by the EC and the European Federation of Pharmaceutical Industries and Associations (EFPIA) under the Innovative Medicines Initiative (IMI). It consists of 29 public and private partners coordinated by the European Medicines Agency. The goal of PROTECT is to strengthen the monitoring of benefit-risk of medicines in Europe by developing innovative methods that will enhance the early detection and assessment of adverse drug reactions from different data sources and enable the integration and presentation of data on benefits and risks. A methodological framework for pharmacoepidemiological studies is to be developed and tested to allow data mining, signal detection and evaluation in different types of datasets, including spontaneous reports, registries and other electronic databases.

Both the US and EU initiatives aim at developing methods for better surveillance of medicinal products, in view of the shortcomings of the current pharmacovigilance system. This deliverable describes the methods of signal detection that have been tested in EU-ADR.

What is a signal?

Key attributes of a signal in pharmacovigilance[17]:

- It is based on information from one or more sources (including observations and experiments), suggesting an association (either adverse or beneficial) between a drug or intervention and an event or set of related events (e.g. a syndrome).
- It represents an association that is new and important, or a new aspect of a known association, and has not been previously investigated and refuted.
- It demands investigation, being judged to be of sufficient likelihood to justify verification and, when necessary, remedial actions.

Overview of this report

We will first provide an overview of methods, including methods that were developed within the EU-ADR project. Methods development was distributed across UNIMIB, LSHTM, PHARMO and EMC.

Subsequently we describe the software architecture used for executing these methods in a distributed database environment, which includes the Jerboa tool created within EU-ADR. For evaluation of the methods, a reference set was developed containing known true adverse drug reactions and drugs and potentially adverse events that are believed to not be related. This reference set was used to evaluate the ability of the various methods in distinguishing true signals from false signals.

3. Overview of methods

Detection of drug safety signals has traditionally been carried out by a systematic manual expert review of spontaneous reports sent by physicians and registered in pharmacovigilance database systems. However, qualitative review of all reported drug-adverse event combinations has become increasingly difficult and impractical because of the constant increase in the number of cases and the continuous development of new drugs. Remarkable though the human brain is as an instrument for seeing patterns, on the scale of hundreds of thousands of reports the requirement of memory is excessive and sheer volume hinders elucidation of possible informative comparisons[18]. To address the difficulties in recognizing patterns in large volumes of data during the last years quantitative signal detection methods have been developed to supplement qualitative clinical methods. While these automated methods cannot replace expert clinical reviewers, they can perform a preliminary sifting of enormous quantities of information, systematically discarding the many comparisons that pose no problems and drawing attention to the few that appear to need more detailed study.

Commonly used automated quantitative methods include data mining techniques that search databases for significant occurrence disproportionalities. Dependencies between drug-adverse event pairs are based on an underlying model of statistical association[19]. These methods include Bayesian methods, proportionate reporting ratios and reporting odds ratios, all of which have been implemented in databases of spontaneous reports of suspected drug-related adverse reactions.

Signal detection methods in EU-ADR

In EU-ADR we use various signal detection methods, including methods based on existing methods used in traditional spontaneous databases, methods from the area of epidemiology, and methods specifically developed for observational data. **Table 1** lists the methods which are currently being used. We are continuously developing and testing new methods in collaboration with the OMOP consortium.

Signal detection on electronic health records potentially may augment the number of the potential signals, in particular since the exposure/disease association is calculated statistically without human interpretation about the plausibility and alternative explanations. In pharmacoepidemiology typical alternative explanations for drug-event associations are bias and confounding. It is the view of EU-ADR that automated methods to detect or look at these alternative explanations should be part of the signal detection methods. Confounding (a false association is found because a third factor is associated with exposure and an independent risk factor for the outcome) is very likely in pharmacoepidemiology and in particular confounding by indication. Confounding can be addressed by randomization, matching, restriction and adjustments. Randomization is the allocation of patients to either the exposure group or control group, and is typically done in clinical trials but impossible in retrospective observational studies. Matching is the practice of only comparing patients that have similar characteristics such as age and sex. Restriction refers to restricting to a homogeneous subgroup of the population, and adjustment refers to including potentially confouding variables in the analysis. In EU-

ADR we primarily choose for matching and adjustments to look at the effect of confounding. This is mostly done in the case based methods.

Bias occurs due to errors in measuring exposure or outcomes, and may have various reasons such as the wrong definition of the exposure window of relevance, misclassification of the outcome (case is not a true case), wrong assessment of the timing of disease onset which may lead to protopathic bias (e.g. drugs being prescribed for the symptoms of the disease which then may seem to be associated).

Table 1. Overview of the methods currently used in EU-ADR. * Note: even though the BHM was currently only applied to the IRR, it can be applied to other types of estimates as well.

	SRS methods	**Cohort methods**	**Case-based methods**
Frequentist	Proportional Reporting Ratio (**PPR**)	Incidence Rate Ratio (**IRR**)	Matched case-control (**CC**)
	Reporting Odds Ratio (**ROR**)		Self-Controlled Case Series (**SCCS**)
Bayesian	Gamma Poisson Shrinker (**GPS**)	Longitudinal GPS (**LGPS**)	
	Bayesian Confidence Propagation Neural Network (**BCPNN**)	Bayesian Hierarchical Model (**BHM**)*	
Elimination of protopathic bias	**LEOPARD**		

Spontaneous Reporting System methods

Signal detection methods originally developed for SRSs can also be used on EHR data by transforming the data to a format suitable for these methods. For this transformation we assume that whenever one of the events of interests occurs during a period associated with the exposure to a drug, that this will lead to a 'report' describing a potential ADR involving the drug and event. The number of report for a particular drug-event pair can then be used as if it is a report count from a spontaneous reporting system, as shown in **Table 2**. In this table, w_{00} is the number of events A that occurred during exposure to drug X, w_{01} is the number of events of a different type than A that occurred during exposure to X, w_{10} is the number of events A that occurred during exposure to drugs other than X, and w_{11} is the number of events of a different type than A, that occurred during exposure to drugs other than X.

Based on this table, the different metrics described below can be calculated. The key disadvantages of these disproportionality methods are that they do not use all the information that is available in the longitudinal health records but focus only on the cases, and it is difficult to adjust for confounding factors. They can be regarded as easy screening methods and can be scaled easily to large healthcare databases as they are not computationally intensive. The LGPS method was developed in EU-ADR to take exposure time into account for estimation. The following disproportionality methods were included in the EU-ADR evaluation:

Table 2. Overview of report counts generated for event A and drug X.

	Event A	Not Event A
Drug X	w_{00}	w_{01}
Not drug X	w_{10}	w_{11}

- **Proportional Reporting Ratio (PRR)** is the ratio of the proportion of all reported cases of the event of interest among people exposed to a particular drug compared with the corresponding proportion among people exposed to all drugs[20].

- **Reporting Odds Ratio (ROR)** is the reformulation of the PRR as an odds ratio [21].

- **Gamma Poisson Shrinker (GPS)** also determines the disproportionality of reports for a particular drug compared to all exposure, but uses an empirical Bayesian model to shrink relative risk estimates when less data is available[22].

- **Bayesian Confidence Propagation Neural Network (BCPNN)** works similarly to GPS, in that it also uses a Bayesian model to shrink estimates of risk. Typically, the output of a BCPNN is expressed as the Information Component (IC); the logarithm of the ratio between observed and expected number of reports for a particular drug-event pair[23].

Cohort methods

One of the limitations of the SRSs and their methods is that only numerator data is available: the number of people on drugs that have events. What is missing is the denominator data: the number of people that are exposed to the drugs. In longitudinal databases this information is readily available, as well as the length of exposure, and this information is used in the cohort methods

- **Incidence Rate Ratio (IRR)** is calculated as the ratio between the incidence rate during exposure to the drug compared to a background incidence rate. A Mantel-Haenszel test is used to test the differences between the incidence rates, typically correcting for age and gender. Important parameter settings are the length and calculation of the exposure window. The default setting in EU-ADR is the legend duration that is supplied by the database holders. In some

databases this is based on the prescribed duration (IPCI, PHARMO, QRESEARCH, PEDIANET) in other databases this is based on the defined daily dose and the quantity prescribed (Lombardy, Aarhus, Health Search, ARS). Events are assigned to exposure if they occur during the exposure duration, there is no carry over period or lag time in the default settings.

- **Longitudinal GPS (LGPS)** is an adaptation of the GPS to longitudinal data, and was developed in the EU-ADR project[1].

- **Bayesian Hierarchical Model (BHM)** uses a full Bayesian approach to perform shrinkage, but instead of using a single prior distribution for all drugs, priors are also created for classes and super classes of drugs. The BHM combines statistical models for the observations given the parameters (likelihood) and the parameters themselves (priors). In the application reported here, the incidence rate is modeled using a Poisson process, and the priors as a hierarchy (using guidance from Gelman[24]). The groupings forming the hierarchy are decided a priori based on criteria of similarity between drugs; in this case we have used ATC coding levels based on organ/systems and therapeutic or chemical characteristics. Berry and Berry[25] used a similar hierarchical approach, with a hierarchy based on related outcomes rather than drugs. The BHM shrinks the original 'frequentist' estimates to give an updated posterior distribution of each individual drug to the group mean and reduces its variance. This is because the posterior considers both the data provided by the drug and by the other drugs in the same group. Shrinkage is stronger for drugs with an initial large variance (less information) and larger effects. These novel methods can offer key advantages by reducing the likelihood of false positive or false negative results obtained from the data. The approach offers further opportunities beyond the application reported here. More specific comparisons are possible, for example between related drugs from the same group or perhaps higher level groups, and potential extensions include using regression methods to adjust for co-prescriptions of drugs[26, 27]. Although the BHM is grouped here with the cohort methods, it can also be applied to other types of relative risk estimates.

Case based methods

Several analytical epidemiological methods start with the diseased persons (cases) and compare these with a sample of the population that gives rise to the cases (i.e. the controls) to evaluate differences in exposure status. Since the case based methods are more efficient in terms of data needs (exposure assessed only at one point in time), they allow for easier adjustments for confounding factors either through matching or statistical adjustments.

- **Case Control (CC)** starts with all cases (i.e. subjects who had a particular event of interest), and finds for every case a predefined number of controls (in our experiments two controls per

case), where controls should have the same characteristics as the cases, such as age and gender. For both cases and controls, the exposure to drugs is determined at the time of event (also known as the index date). A conditional logistic regression is used to determine the effect size of exposure to a drug. In our experiments, one covariate was used: the **drug count**, which is the number of different drugs the subject was exposed to in one year prior to the event date, until one month prior to the event date. The drug count is assumed to be an indication of overall patient health, and was included in the logistic regression.

- **Self-Controlled Case Series (SCCS)** investigates the association between acute outcomes and transient exposures, whereby cases are used as their own controls. In essence, the SCCS is a Poisson regression conditioned on the patient[28]. Only information of cases is used in this analysis, all other persons are ignored.

Other methods

One other method not categorized elsewhere remains:

- **Longitudinal Evaluation of Observational Profiles of Adverse events Related to Drugs (LEOPARD)** attempts to detect protopathic bias. For every suspect drug-event combination, the number of prescription starting in a 51 day window around the event date are counted, as shown in **Figure 1** for the drugs Pantoprazole and Naproxen around the event 'upper gastro-intestinal bleeding' (data from the IPCI database). The number of prescriptions in the 25 days prior to the event is compared to the number or prescriptions starting in the 25 days after the event. If the number of prescriptions increases after the event date, this is an indication that the drug is used to treat the event or a precursor of the event, rather than cause it. This is tested using a binomial test. In the examples of figure 1, Pantoprazole has $p < 0.001$, indicating the signal is probably caused by protopathic bias, whilst Naproxen has $p = 1.00$, indicating that the signal is probably not caused by protopathic bias. A signal is considered to be caused by protopathic bias if the p-value is below 0.5. LEOPARD was developed in the EU-ADR project[1].

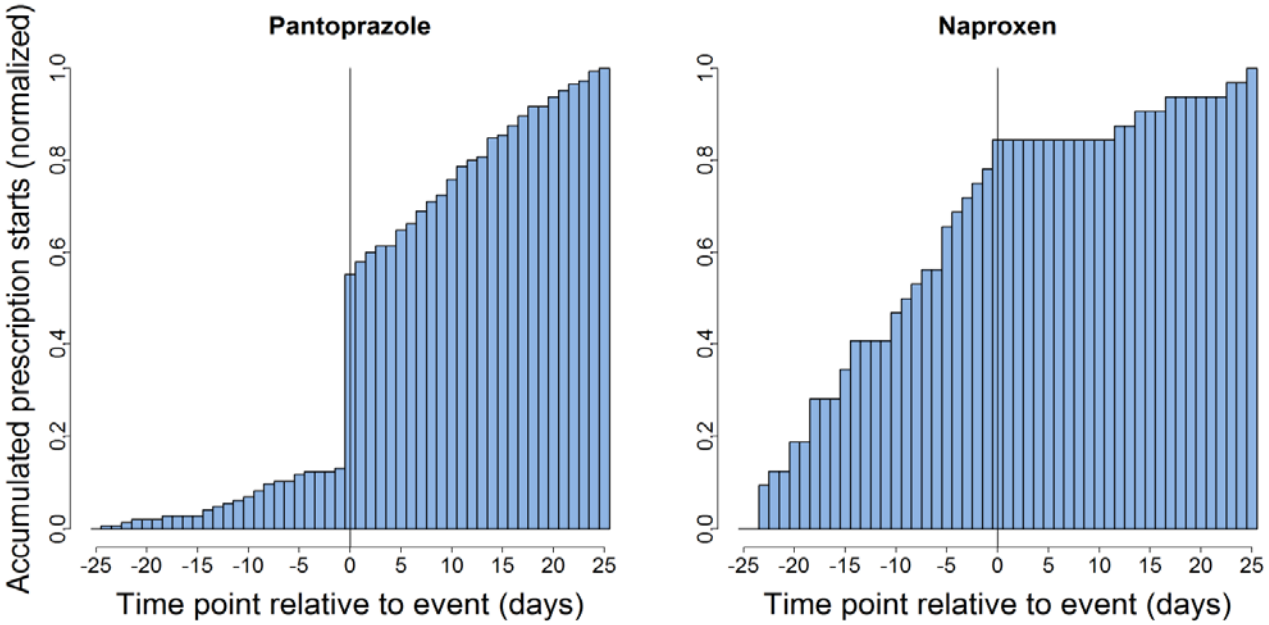


Figure 1. Empirical cumulative distribution functions of prescription starts in a window around upper gastro-intestinal bleeding occurrences for Pantoprazole and Naproxen.

4. Software architecture

Figure 2 shows an overview of the software architecture used in EU-ADR[29].

From each database, data is extracted and stored in three flat text files containing information about prescriptions, events, and patients, respectively. The event file is populated with potentially adverse events defined in the event harmonization process.

The three input files are read by a stand-alone application called Jerboa. Jerboa was developed within the EU-ADR project. It is written in Java, and can therefore easily run on any platform. Jerboa is executed by the database owners in their local environment, and transforms the input files into aggregated output files. For each type of analysis a different type of output is generated. The aggregated output formats used for the different methods described in this deliverable are detailed in **Appendix B**. Most importantly, these aggregated data are completely anonymized.

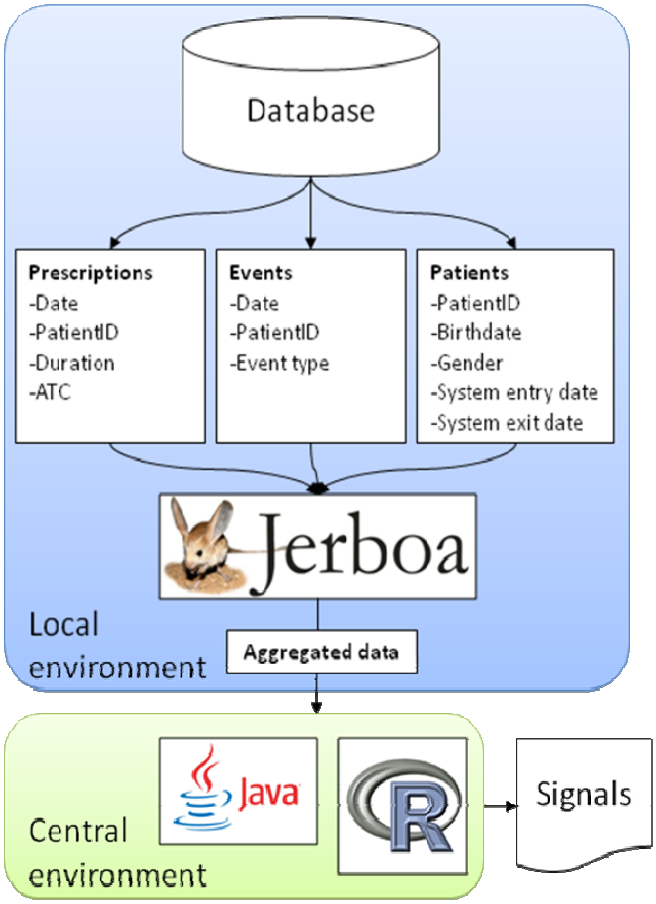


Figure 2. Overview of the software architecture for signal detection

The aggregated data are encrypted and transmitted to the central environment for further processing. Currently the statistical analyses are performed in Java and in R, but other software could also be used. The result is a list of drug-event pairs with estimates of relative risks from each of the methods.

Jerboa was originally developed within EU-ADR, but is now also used in several other European projects such as the SOS project, ARITMO project, and the VAESCO project.

5. Reference set

The performance of the signal detection methods was evaluated by comparing how well the methods could distinguish between known ADRs, and drug-event pairs where the drug is probably not causing the event. The procedure employed in the construction of the reference set is outlined in **Figure 3**. It was first necessary to ensure that the drug-event associations to be included in the reference set could be found in the EU-ADR database network. That is, there should be adequate exposure to the drugs to permit detection of an association with a particular adverse event, if present. In an earlier publication we described the sample size calculations used to derive the total amount of person-years (PYs) of drug exposure required to detect an association between a drug and a particular event over varying magnitudes of relative risk, given pooled population-based incidence rates (IR) estimated directly within the EU-ADR network. A series of steps was subsequently employed to determine which among the remaining drug- event associations (i.e., with adequate exposure to detect an association) were previously known from existing data sources.

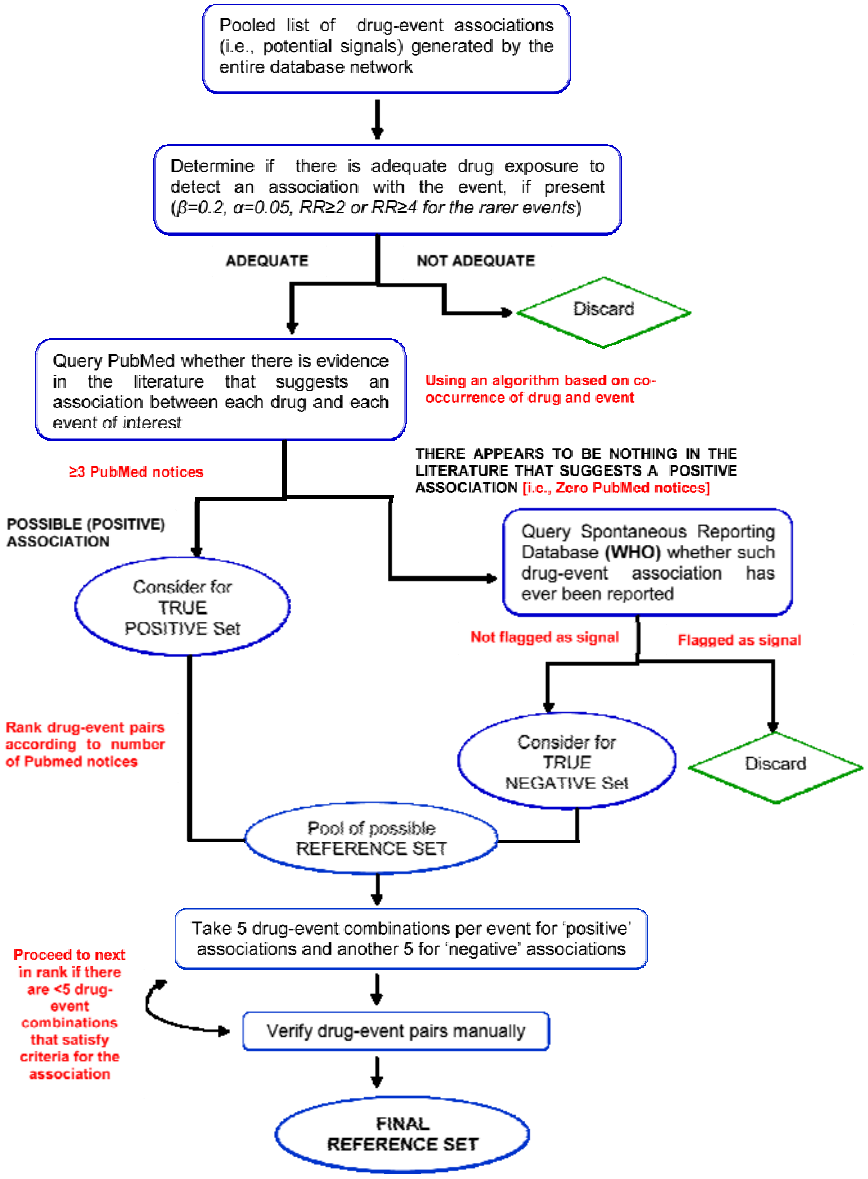


Figure 3. Flowchart showing the process of the construction of the Reference Set.

Retrieving information from published literature

A subset of MEDLINE was downloaded (via PubMed) and imported in a database including all the citations with the "adverse effects" MeSH subheading. For each citation the PubMed identification (PMID), MeSH descriptors, subheadings, substances, and date of creation of the citation were obtained.

Co-occurrences of four elements in a citation were noted: (1) the drug (from "substances" OR "MeSH heading" fields); (2) adverse effect and the two subheadings, 'adverse effect' and 'contraindications'. Drugs from the "substances" field were taken into account only if their pharmacological action was qualified by the subheading "adverse effects." Hence, in this case, the pharmacological action was an additional element that had to be taken into account. This latter requirement was an attempt to ensure that there would be a link between an adverse event and a drug in the context of drug safety and not just a co-occurrence in a MEDLINE citation[30].

Filtering possible known associations

The drug-event pairs were ranked according to the number of PubMed citations with co-occurrence of the drug and the adverse event of interest. For the pool of true positive associations, we considered the drug-event pairs with the highest number of citations. This meant that more investigations were performed – and published - on these drug-adverse event pairs. A drug-event pair was considered for the pool of true negative associations if there were no PubMed citations with co-occurrence of the drug and the adverse event of interest. The pool of true negative drug-event pairs was further verified using the World Health Organization's spontaneous ADR reporting database VigiBase to determine whether any of these associations have previously been flagged as a potential signal[31]. Supplementary information for both positive and negative associations was also obtained from the Summary of Product Characteristics (SPCs) or product labels.

Grading the evidence from literature

Table 3 shows the scheme that was used as guide to evaluate the evidence from literature. Manual verification of 'true positive' and 'true negative' associations was conducted by two physicians with expertise in clinical medicine, epidemiology, and pharmacovigilance. A panel of experts adjudicated equivocal cases as well as any disagreements between evaluators. The following indices of agreement between evaluators were assessed: (1) proportion of overall agreement; (2) proportions of specific agreement; and (3) corresponding kappa statistic, κ, unweighted and weighted using quadratic weighting.

Table 3. Levels of evidence for evaluating drug safety information in the literature

Level of Evidence	**Description**
Level I	Evidence from at least one (properly designed) randomized controlled trial or meta-analysis.
Level II	Evidence from at least one observational study (cohort/case–control/case-cross over/self-controlled case series) OR from at least three (3) published case reports from different sources and concerning different patients.
Level III	Evidence from not more than two (2) published case reports OR from unpublished reports in pharmacovigilance databases and no further publications on the potential ADR in the literature.
Level IV	Included in drug label (Summary of Product Characteristics, SPC) but no case reports or published studies.
Level V	No evidence from published literature or from WHO spontaneous reporting database and not mentioned in SPC.
Recommendations: Levels I and II → TRUE POSITIVE Levels III and IV → CANNOT BE DETERMINED → DISREGARD Level V → TRUE NEGATIVE	

Resulting reference set

94 drug-event combinations comprised the reference set, which included 44 true positive associations and 50 true negative associations for 10 events of interest: bullous eruptions; acute renal failure; anaphylactic shock; acute myocardial infarction; rhabdomyolysis; aplastic anemia; neutropenia; cardiac valve fibrosis; acute liver injury; and upper gastrointestinal bleeding. For cardiac valve fibrosis, there was no drug with adequate exposure in the database network to permit detection of a true positive association. The complete set can be found in **Appendix A.**

6. Method of comparison

The reference set was used to evaluate the performance of each of the methods, using seven of the eight databases in the EU-ADR project.

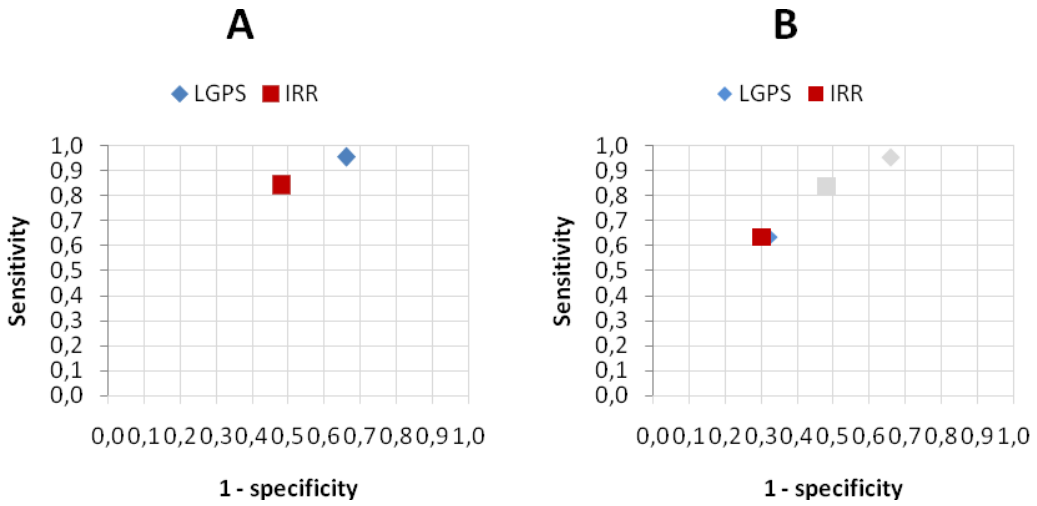
Performance metrics

Typically, the output of a signal detection method is turned into a binary decision (positive or negative) by employing a threshold, for instance on the p value of the test whether the relative risk is unequal to 1. A commonly used but completely arbitrary threshold is $p < 0.05$. Drug-event pairs for which the p value is below the threshold are called positive signals, the others are called negative signals. By comparing the outcome of the method to the reference set, each drug-event pair can be classified into one of four categories as shown in **Figure 4**. The performance of the method on the entire reference set can be summarized by two statistics: sensitivity = TP/(TP+FN), and specificity = TN/(TN+FP).

		Method: Positive	Method: Negative
Reference set	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Figure 4. Comparison of method outcome to reference set.

The sensitivity and specificity of two methods, using a cutoff of $p < 0.05$ is shown in **Figure 5A**. From this graph, one could conclude for example that IRR is a better method, because it has a higher specificity than LGPS with just a small decrease in sensitivity. However, suppose we want to make the definition of a positive signal stricter, for instance for correcting for multiple testing. **Figure 5B** shows the sensitivity and specificity using a cutoff of $p < 0.001$. Not only does LGPS now have a much higher specificity than the original IRR score, the difference between both methods has become very small.



Figures 5A and 5B. Sensitivity and 1-specificity of the LGPS and IRR methods with threshold A: $p < 0.05$, and B: $p < 0.001$.

As the example above shows, sensitivity and specificity can be exchanged by picking a different threshold. Comparing single sensitivities and specificities is therefore not informative. Typically for method comparison the Receiver Operator Characteristics (ROC) curve is drawn, showing all sensitivities and specificities as shown in **Figure 6**. Such a ROC curve can subsequently be summarized into one statistic: the Area under the Curve (AuC). The AuC indicates the overall performance of a method, independent of any threshold. An AuC of 0.5 indicates random performance, an AuC of 1.0 indicates a perfect performance. The AuC has been used as primary performance indicator.

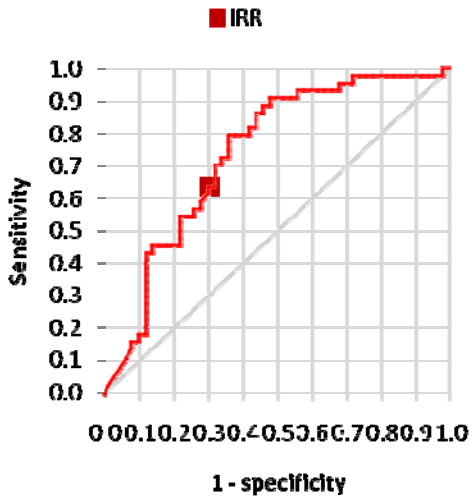


Figure 6. Receiver Operator Characteristics curve for IRR.

Ranking criteria

For each of the signal detection methods, a measure was selected for ranking drug-event pairs from most probable ADR to least likely. This ranking was the basis for the ROC curve and AuC performance metric. The ranking criteria are specified in **Table 4.**

Table 4. Criteria used for ranking drug-event pairs according to the different methods.

Disproportionality methods	**Ranking criteria**
Proportional Reporting Ratio (**PPR**)	PRR
Reporting Odds Ratio (**ROR**)	ROR
Gamma Poisson Shrinker (**GPS**)	Point estimate of the RR
Bayesian Confidence Propagation Neural Network (**BCPNN**)	Information Component (IC)
Incidence Rate Ratio (**IRR**)	IRR
Longitudinal GPS (**LGPS**)	Point estimate of the IRR
Bayesian Hierarchical Model (**BHM**)	Point estimate of the IRR
Matched case-control (**CC**)	Relative risk (estimate of the beta coefficient in the conditional logistic regression)
Self-Controlled Case Series (**SCCS**)	Relative risk (estimate of the beta coefficient in the conditional poisson regression)

Common settings

For all methods, these specifications were used to define exposures and outcomes:

- **Incident events.** Only the first occurrence of an event was considered. Patient time after an event was completely ignored. The main reason for this is that in EHR data it is often difficult to distinguish between a recurrence of an event, or whether a reference is made to the event that occurred earlier.
- **Run-in period of 365 days**. In order to determine that an event is incident, some patient time has to be available before the event occurred. Hence, during the first year of observation subjects were not considered for events or exposure counts, but events during this so-called run-in period were used to determine whether later events were truly incident events. This run-in period was not used for children younger than one year at the start of observation.

- **Exposure window definition**. Exposure to a drug was defined as the duration of the prescription, excluding the first day of the prescription. If two prescriptions of the same drug overlapped in time, the exposure was assumed to start the day after the first prescription of the first prescription, and end on the last day of the last prescription.
- **Age stratification**. Whenever appropriate, age was stratified in 5-year age ranges.
- **Independence of drug risks**. Currently, every drug-event pair is evaluated separately. Co-medication is not taken into account.

LEOPARD was considered to be potentially complimentary to all methods, and was therefore applied as a filter to the output of each method. LEOPARD can be applied at the level of the individual drug, but it can also be applied to a group of drugs. By grouping drugs with the same 4 higher level ATC digits (i.e. drugs with the same indication), LEOPARD has proven more able to pick up protopathic bias. Signals that are flagged by LEOPARD either at individual or at group level were ranked lower in the list of signals than signals that were not flagged when calculating the AuC.

Combination of databases

The information of the different databases has to be combined to generate a single score per drug-event pair, per method. In principle there are two approaches: Pooling of the data as if the databases together form one large database, or computing the score per database and using meta-analysis techniques to combine the scores. We have tested both pooling of data, and meta-analysis assuming random effects.

7. Results of the method comparison

Figure 7 shows the amount of data, measured in patient time, available from each database over time. In total, 146,830,906 patient years of 20,042,652 subjects were included in the study.

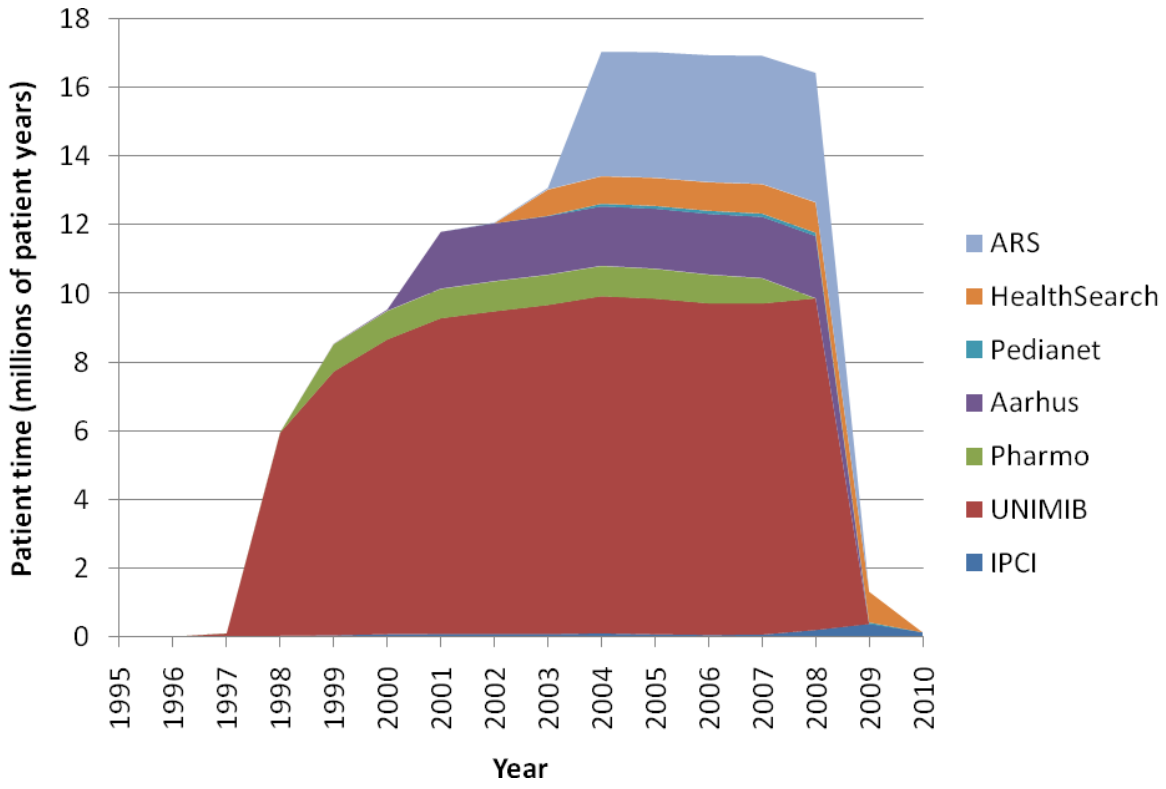


Figure 7: Distribution of patient data per database over time.

Overall performance of methods

Figure 8 and **Figure 9** show the performance of the different methods on the reference set, using meta-analysis for random effects, and data pooling, respectively.

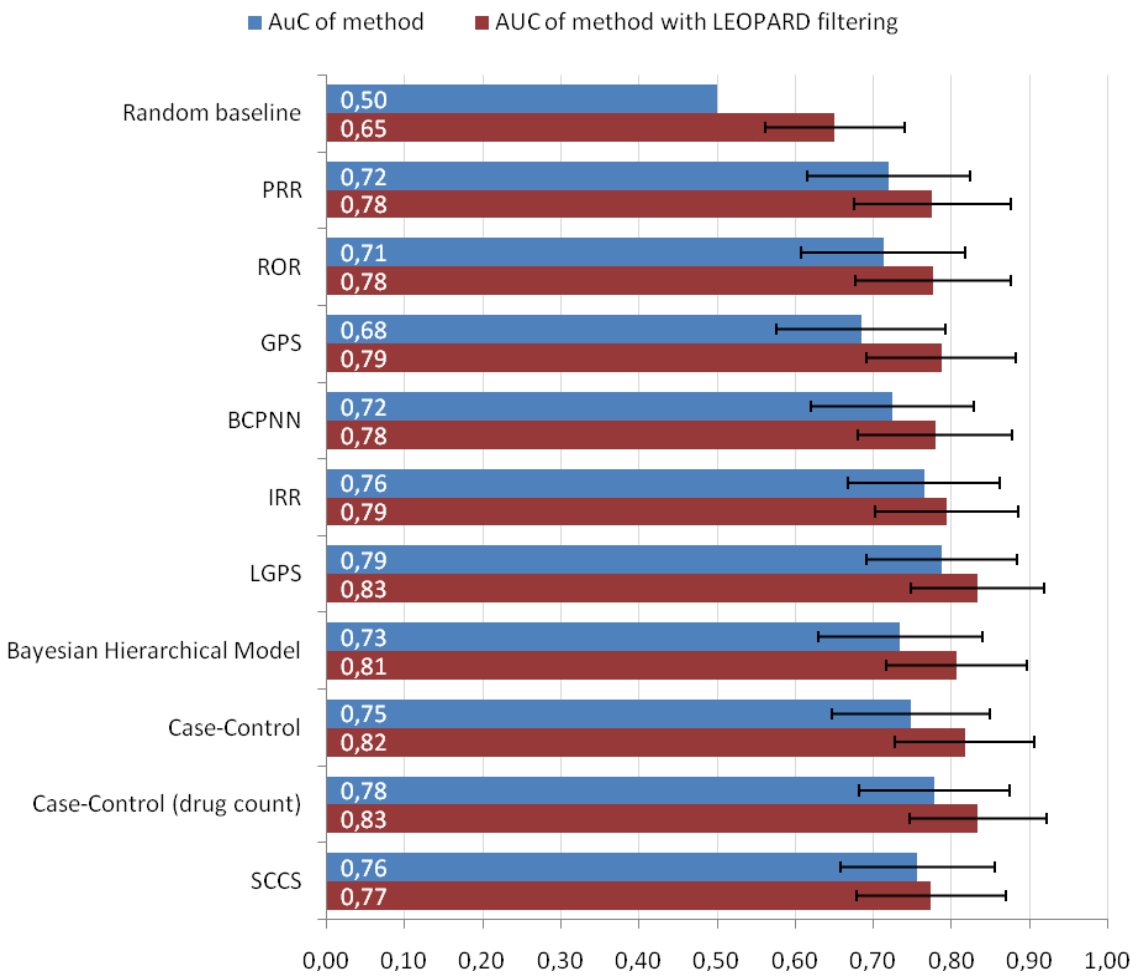


Figure 8. Area under the ROC curve for all methods, with and without LEOPARD filtering. Combination across databases was performed using meta-analysis for random effects. Error bars indicate 95% confidence interval.

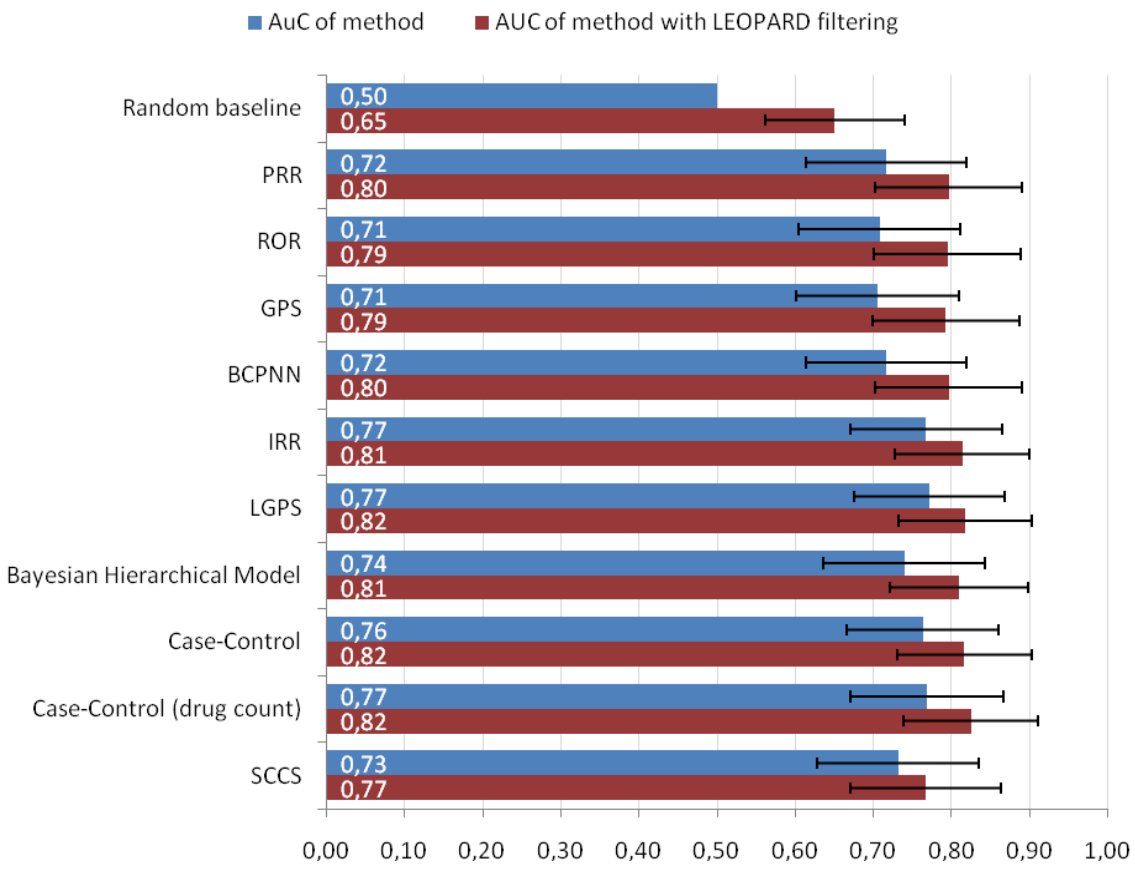


Figure 9. Area under the ROC curve for all methods, with and without LEOPARD filtering. Combination across databases was performed by pooling data. Error bars indicate 95% confidence interval.

These figures show that all methods perform better than random baseline, that the LEOPARD filtering for protopathic bias always improves performance, but less so for methods that are already performing well, and that performance of methods does not differ that much. In general LGPS and case-control adjusting for drug count seem to slightly outperform the other methods, although this is certainly not statistically significant.

Figure 10 shows the relative risk estimates of the LGPS method on the reference set, and includes the LEOPARD filtering. If we were to use a threshold value of $p < 0.05$, in other words, require the 95%

confidence interval to be above 1, and were to remove signals flagged by LEOPARD, this combination of methods would achieve a sensitivity of 0.80, and a specificity of 0.70.

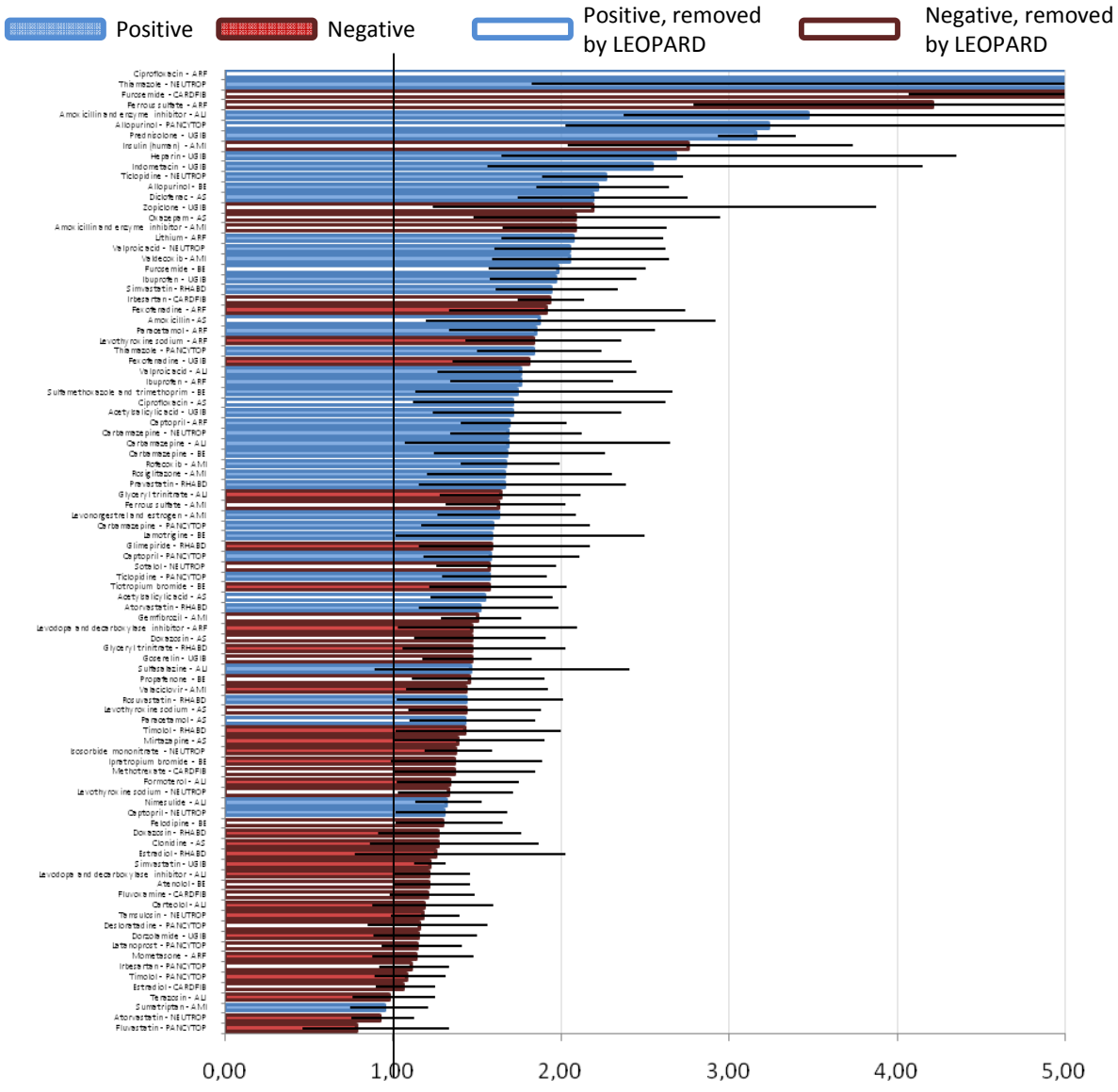


Figure 10. Relative risk estimates of the LGPS method on the reference set, using meta-analysis for random effects. Error bars indicate 95% confidence intervals.

8. Conclusions

In general the performance of methods is high, with the best performing method achieving an area under the ROC curve of 0.83. When using a default threshold of p<.05, a sensitivity and specificity of 0.80 and 0.70, respectively was achieved. On the one hand, this is not surprising, since the reference set was limited to drugs with a large amount of exposure in the different databases. This probably explains why the Bayesian methods are not performing much better than frequentist methods, as these methods are designed to deal with sparse data. On the other hand, the high performance is surprising given the fact that most of these methods are very simple: the best performing methods do not take any covariates into account, other than age and gender, and compare one drug to one event of interest at a time. It is expected that by including covariate data, performance could be improved further, but the data needed for this is not easily available in all databases. For example: diabetes is a risk factor for myocardial infarction, and diabetics will tend to be exposed to anti-diabetic drugs. Diabetes is therefore a confounder between myocardial infarction and anti-diabetics, and should ideally be included in the analysis. However, diabetes will be coded differently in the different databases in EU-ADR, and time-consuming harmonization would be needed to extract this data in a uniform way. Since for every drug-event combination there can be different potential confounders, many such variables would need to be extracted, and this currently not feasible.

The software architecture, centered around the Jerboa tool, allows for incorporation of data from different databases without sacrificing privacy and database owner sovereignty.

The next step within the EU-ADR project will be to evaluate the sensitivity of the different methods to different parameter settings. Furthermore, all methods for signal detection are focused primarily on acute or short term ADRs whilst many important ADRs such as cancer only occur after prolonged exposure, and will also not be diagnosed until the disease has progressed a while. Future research will focus on methods that can detect these long term ADRs using observational data. Another future direction of research could be the development of techniques for adjusting for confounding without the use of harmonized co-variates, for instance by using summary statistics such as propensity scores instead.

The development and evaluation of techniques as described here does not guarantee improved drug safety in Europe. Similar to the Sentinal initiative in the US, it is advisable that the knowledge and technology developed in this project is transferred from the research domain into actual application, and continuous monitoring.

References

1. Schuemie, M.J. Methods for drug safety signal detection in longitudinal observational databases: LGPS and LEOPARD. *Pharmacoepidemiol Drug Saf* **20**, 292-299 (2011).
2. Mann, R. & Andrews, E. (eds.) Pharmacovigilance. (John Wiley & Sons, 2002).
3. Ahmad, S.R. Adverse drug event monitoring at the Food and Drug Administration. *J Gen Intern Med* **18**, 57-60 (2003).
4. Olsson, S. The role of the WHO programme on International Drug Monitoring in coordinating worldwide drug safety efforts. *Drug Saf* **19**, 1-10 (1998).
5. Avorn, J. Evaluating drug effects in the post-Vioxx world: there must be a better way. *Circulation* **113**, 2173-2176 (2006).
6. Rodriguez, E.M., Staffa, J.A. & Graham, D.J. The role of databases in drug postmarketing surveillance. *Pharmacoepidemiol Drug Saf* **10**, 407-410 (2001).
7. Meyboom, R.H., Egberts, A.C., Edwards, I.R., Hekster, Y.A., de Koning, F.H. & Gribnau, F.W. Principles of signal detection in pharmacovigilance. *Drug Saf* **16**, 355-365 (1997).
8. Belton, K.J. Attitude survey of adverse drug-reaction reporting by health care professionals across the European Union. The European Pharmacovigilance Research Group. *Eur J Clin Pharmacol* **52**, 423-427 (1997).
9. Alvarez-Requejo, A., Carvajal, A., Begaud, B., Moride, Y., Vega, T. & Arias, L.H. Under-reporting of adverse drug reactions. Estimate based on a spontaneous reporting scheme and a sentinel system. *Eur J Clin Pharmacol* **54**, 483-488 (1998).
10. Eland, I.A., Belton, K.J., van Grootheest, A.C., Meiners, A.P., Rawlins, M.D. & Stricker, B.H. Attitudinal survey of voluntary reporting of adverse drug reactions. *Br J Clin Pharmacol* **48**, 623-627 (1999).
11. De Bruin, M.L., van Puijenbroek, E.P., Egberts, A.C., Hoes, A.W. & Leufkens, H.G. Non-sedating antihistamine drugs and cardiac arrhythmias -- biased risk estimates from spontaneous reporting systems? *Br J Clin Pharmacol* **53**, 370-374 (2002).
12. Meyboom, R.H., Hekster, Y.A., Egberts, A.C., Gribnau, F.W. & Edwards, I.R. Causal or casual? The role of causality assessment in pharmacovigilance. *Drug Saf* **17**, 374-389 (1997).
13. Ash, J.S. & Bates, D.W. Factors and forces affecting EHR system adoption: report of a 2004 ACMI discussion. *J Am Med Inform Assoc* **12**, 8-12 (2005).
14. Schade, C.P., Sullivan, F.M., de Lusignan, S. & Madeley, J. e-Prescribing, efficiency, quality: lessons from the computerization of UK family practice. *J Am Med Inform Assoc* **13**, 470-475 (2006).
15. McClellan, M. Drug safety reform at the FDA--pendulum swing or systematic improvement? *N Engl J Med* **356**, 1700-1702 (2007).
16. Platt, R. Challenges for the FDA: The Future of Drug Safety, Workshop Summary., Edn. 2011/04/01. (National Academy of Sciences, 2007).

17. Hauben, M. & Aronson, J.K. Defining 'signal' and its subtypes in pharmacovigilance based on a systematic review of previous definitions. *Drug Saf* **32**, 99-110 (2009).
18. Finney, D.J. The detection of adverse reactions to therapeutic drugs. *Stat Med* **1**, 153-161 (1982).
19. Hauben, M. & Zhou, X. Quantitative methods in pharmacovigilance: focus on signal detection. *Drug Saf* **26**, 159-186 (2003).
20. Evans, S.J., Waller, P.C. & Davis, S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol Drug Saf* **10**, 483-486 (2001).
21. Rothman, K.J., Lanes, S. & Sacks, S.T. The reporting odds ratio and its advantages over the proportional reporting ratio. *Pharmacoepidemiol Drug Saf* **13**, 519-523 (2004).
22. DuMouchel, W. Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System. *The American Statistician* **53**, 177-190 (1999).
23. Norén, G.N., Bate, A., Orre, R. & Edwards, I.R. Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events. *Statistics in Medicine* **25**, 3740-3757 (2006).
24. Gelman, A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**, 515--534 (2006).
25. Berry, S.M. & Berry, D.A. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. *Biometrics* **60**, 418-426 (2004).
26. Caster, O., Norén, G.N., Madigan, D. & Bate, A. Large-scale regression-based pattern discovery: The example of screening the WHO global drug safety database. *Statistical Analysis and Data Mining* **3**, 197-208 (2010).
27. Madigan, D., Patrick Ryan, P., Simpson, S., and Zorych, I in Bayesian Statistics 9. (ed. J. Bernardo, Bayarri, M) (OUP, 2010).
28. Whitaker, H.J., Farrington, C.P., Spiessens, B. & Musonda, P. Tutorial in biostatistics: the self-controlled case series method. *Stat Med* **25**, 1768-1797 (2006).
29. Coloma, P.M., Schuemie, M.J., Trifiro, G. et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiol Drug Saf* **20**, 1-11 (2010).
30. Avillach P, D.J., Diallo G, Joubert M, Thiessard F, Mougin F, Trifirò G, Fourrier-Réglat A, Pariente A, Fieschi M. Design and Validation of an Automated Method to Detect Known Adverse Drug Reactions in MEDLINE: a Contribution to the European EU-ADR Project. . *to be published*.
31. Trifiro, G., Patadia, V., Schuemie, M.J. et al. EU-ADR healthcare database network vs. spontaneous reporting system database: preliminary comparison of signal detection. *Stud Health Technol Inform* **166**, 25-30 (2011).

Appendix A: Reference set

Table A.1. Reference Set of True Positive Associations

Event	True Positive Associations	
	ATC	**Name**
Acute Liver Injury (ALI)	N03AF01	Carbamazepine
	N03AG01	Valproic acid
	M01AX17	Nimesulide
	J01CR02	Amoxicillin and clavulanic acid
	A07EC01	Sulfasalazine
Acute Myocardial Infarction (AMI)	M01AH02	Rofecoxib
	A10BG02	Rosiglitazone
	G03AA07	Levonorgestrel and estrogen
	N02CC01	Sumatriptan
	M01AH03	Valdecoxib
Acute Renal Failure (ARF)	C09AA01	Captopril
	M01AE01	Ibuprofen
	N02BE01	Paracetamol
	J01MA02	Ciprofloxacin
	N05AN01	Lithium
Anaphylactic Shock (AS)	B01AC06	Acetylsalicylic acid
	N02BE01	Paracetamol
	J01CA04	Amoxicillin
	J01MA02	Ciprofloxacin
	M01AB05	Diclofenac
Bullous Eruptions (BE)	N03AF01	Carbamazepine
	J01EE01	Sulfamethoxazole and trimethoprim
	N03AX09	Lamotrigine
	M04AA01	Allopurinol
	C03CA01	Furosemide

Cardiac Valve Fibrosis (CARDFIB)	No drug with sufficient exposure that satisfies criteria for True Positive	
Neutropenia (NEUTROP)	H03BB02	Thiamazole
	B01AC05	Ticlopidine
	C09AA01	Captopril
	N03AF01	Carbamazepine
	N03AG01	Valproic acid
Aplastic anemia/ Pancytopenia (AA)	B01AC05	Ticlopidine
	N03AF01	Carbamazepine
	H03BB02	Thiamazole
	M04AA01	Allopurinol
	C09AA01	Captopril
Rhabdomyolysis (RHABD)	C10AA07	Rosuvastatin
	C10AA05	Atorvastatin
	C10AA03	Pravastatin
	C10AA01	Simvastatin
Upper Gastrointestinal Bleeding (UGIB)	N02BA01/B01AC06	Acetylsalicylic acid
	M01AB01	Indometacin
	B01AB01	Heparin
	H02AB06	Prednisolone
	M01AE01	Ibuprofen

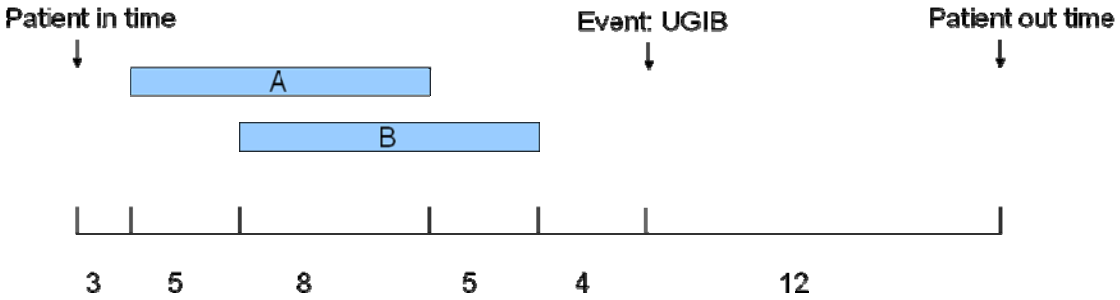
Table A.2. Reference Set of True Negative Associations

Event	ATC	Name
Acute Liver Injury (ALI)	R03AC13	Formoterol
	S01ED05	Carteolol
	G04CA03	Terazosin
	N04BA02	Levodopa and decarboxylase inhibitor
	C01DA02	Glyceryl trinitrate
Acute Myocardial Infarction (AMI)	A10AD01	Insulin (human)
	B03AA07	Ferrous sulfate
	J01CR02	Amoxicillin and clavulanic acid
	J05AB11	Valaciclovir
	C10AB04	Gemfibrozil
Acute Renal Failure (ARF)	R01AD09	Mometasone
	H03AA01	Levothyroxine sodium
	R06AX26	Fexofenadine
	N04BA02	Levodopa and decarboxylase inhibitor
	B03AA07	Ferrous sulfate
Anaphylactic Shock (AS)	N06AX11	Mirtazapine
	H03AA01	Levothyroxine sodium
	C02AC01	Clonidine
	C02CA04	Doxazosin
	N05BA04	Oxazepam
Bullous Eruptions (BE)	C01BC03	Propafenone
	C07AB03	Atenolol
	R03BB01	Ipratropium bromide
	R03BB04	Tiotropium bromide
	C08CA02	Felodipine
Cardiac Valve Fibrosis (CARDFIB)	N06AB08	Fluvoxamine
	L04AX03	Methotrexate
	C09CA04	Irbesartan

	C03CA01	Furosemide
	G03CA03	Estradiol
Neutropenia (NEUTROP)	C07AA07	Sotalol
	H03AA01	Levothyroxine sodium
	C10AA05	Atorvastatin
	C10DA14	Isosorbide mononitrate
	G04CA02	Tamsulosin
Aplastic anemia/ Pancytopenia (AA)	C09CA04	Irbesartan
	C10AA04	Fluvastatin
	S01EE01	Latanoprost
	S01ED01	Timolol
	R06AX27	Desloratadine
Rhabdomyolysis (RHABD)	G03CA03	Estradiol
	C02CA04	Doxazosin
	A10BB12	Glimepiride
	S01ED01	Timolol
	C01DA02	Glyceryl trinitrate
Upper Gastrointestinal Bleeding (UGIB)	R06AX26	Fexofenadine
	C10AA01	Simvastatin
	S01EC03	Dorzolamide
	L02AE03	Goserelin
	N05CF01	Zopiclone

Appendix B: Jerboa output formats

AggregateByATC format



Contains non-exclusive patient time: one patient day can be assigned to several rows. Each row contains at most one ATC code. Exposure time is divided into exposure periods (exposure is accumulated over the last year). Time after exposure is marked separately. Time after an event is marked separately. This format is used for all disproportionality and cohort methods.

Columns

ATC — Zero or one ATC code that was used during the time period, including exposure codes. Codes starting with a P indicate the time after exposure. (VS=0-7, S=8-30, M=31-185, L=186-)
Gender — F = Female, M = Male
AgeRange — Age, divided into categories of 5 years
Days — Days summed over all patients
Subjects
Events — Events that occurred during the period. Format: 'EventType:Count'
PrecedingEventTypes — Zero, one or more events that occurred prior to this time

Example data

ATC	AgeRange	Gender	Days	Subjects	Events	PrecedingEventTypes
	25-29	F	3+5+8+5+4	1	UGIB:1	
	25-29	F	12	1		UGIB
A:VS	25-29	F	5+2	1		
A:S	25-29	F	6	1		
A:PVS	25-29	F	5+2	1		
A:PS	25-29	F	2	1	UGIB:1	
A:PS	25-29	F	12	1		UGIB

B:VS	25-29	F	7	1		
B:S	25-29	F	1+5	1		
B:PVS	25-29	F	4	1		
B:PVS	25-29	F	3	1	UGIB:1	
B:PS	25-29	F	12	1		UGIB

CaseControl format

Contains for every case a number of sampled controls (currently 2 per case), matching controls on index date, age(optional), gender(optional), severity (optional, can be measured using the adapted Chronic Disease Score, or the count of different drugs used in the last year). This format is used for the matched case-control methods.

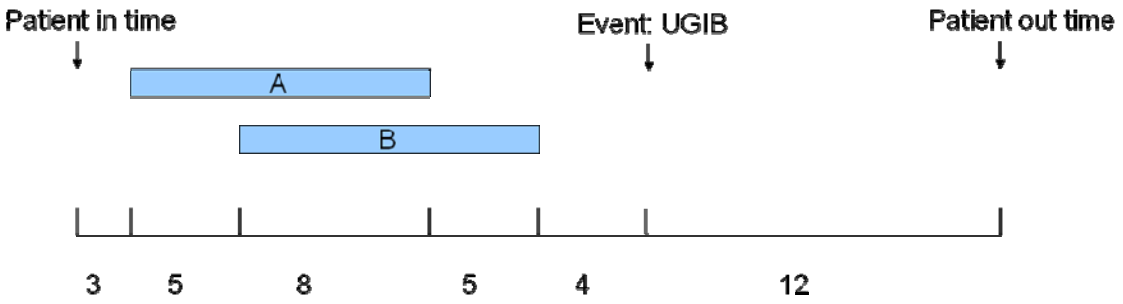
Columns

CaseSetID Identifier for linking cases to matched controls
EventType Type of event that occurred on the index date for the case
isCase Is 1 if the row described the case, 0 if it's a control
Year
Month
Gender F = Female, M = Male
AgeRange Age, divided into categories of 5 years
Current_DaysOfUse ATC codes of drugs used in the month immediately preceding the event (not including the index date itself), including the days of use in that month.
Current_DaysSinceUse ATC codes of drugs used in the month immediately preceding the event (not including the index date itself), including the days between last use and the index date (zero if still being used on the index date).
Past_DaysOfUse Similar to Current_DaysOfUse, but for the period 365 to 31 days before the event
Past_DaysSinceUse Similar to Current_ DaysSinceUse, but for the period 365 to 31 days before the event

Example data

CaseSetID	EventType	IsCase	Year	Month	Gender	AgeRange	current_DaysOfUse	current_DaysSinceUse
0	UGIB	1	2001	4	F	25-29	A:13+B:13	A+9+B:4
0	UGIB	0	2001	4	F	25-29		

SCCS format



For the SCCS, the patient time of all patients with at least one event is split into periods of equal exposure. All these periods are described in the output file. This format is used for the SCCS method.

Columns

PatientID	Unique patient identifier
Year	Noise has been added to the year and month boundaries for de-identification
Month	
Duration	Of the time period (days)
ATC	Zero, one or more ATC codes that were used during the time period
Gender	F = Female, M = Male
AgeRange	Age, divided into categories of 5 years
Event s	Zero, one or more events that occurred during the time period

Example data

PatientID	Year	Month	Duration	ATC	Gender	AgeRange	Events
1045	2001	3	3		F	25-29	
1045	2001	3	5	A	F	25-29	
1045	2001	3	8	A+B	F	25-29	
1045	2001	3	5	B	F	25-29	
1045	2001	4	4		F	25-29	UGIB
1045	2001	4	12		F	25-29	

PrescriptionStartProfiles format

Shows for every drug-event combination the number of prescription starts in a time window around the event index date (window size currently -25 to +25 days). Only the first occurrence is considered, recurrent events are ignored. This format is used for LEOPARD.

Columns

ATC One ATC code
EventType Type of event that is considered in this row
Total Event Count Number of times the event occurred (excluding recurring events)
Datapoint (-25 days) Number of prescription starting exactly 25 days before the event
Datapoint (-24 days) Number of prescription starting exactly 24 days before the event
…

Example data

ATC	EventType	Datapoint (-22 days)	Datapoint (-21 days)	Datapoint (-20 days)	Datapoint (-19 days)	Datapoint (-18 days)	…
A	UGIB	1	0	0	0	0	…
B	UGIB	0	0	0	0	0	…