



UPPSALA
UNIVERSITET

UPTEC X 15 024

Examensarbete 30 hp
Juni 2015

Deviating time-to-onset in predictive models

-detecting new adverse effects from medicines

Caroline Wärn



UPPSALA
UNIVERSITET

Degree Project in Bioinformatics

Master's Programme in Molecular Biotechnology Engineering,
Uppsala University School of Engineering

UPTEC X 15 024		Date of issue: 2015-06	
Author Caroline Wörn			
Title Deviating time-to-onset in predictive models - detecting new adverse effects from medicines			
Abstract <p>Identifying previously unknown adverse drug reactions becomes more important as the number of drugs and the extent of their use increases. The aim of this Master's thesis project was to evaluate the performance of a novel approach for highlighting potential adverse drug reactions, also known as signal detection. The approach was based on deviating time-to-onset patterns and was implemented as a two-sample Kolmogorov-Smirnov test for non-vaccine data in the safety report database, Vigibase. The method was outperformed by both disproportionality analysis and the multivariate predictive model vigiRank. Performance estimates indicate that deviating time-to-onset patterns is not a suitable approach for signal detection for non-vaccine data in Vigibase.</p>			
Keywords <p>Data mining, Kolmogorov-Smirnov test, predictive model, signal detection, time-to-onset</p>			
Supervisors Ola Caster Uppsala Monitoring Centre			
Scientific reviewer Mats Gustafsson Uppsala University			
Project name -		Sponsors -	
Language English		Security -	
ISSN 1401-2138		Classification -	
Supplementary bibliographical information -		Pages 54	
Biology Education Centre Box 592, S-751 24 Uppsala		Biomedical Center Tel +46 (0)18 4710000	
		Husargatan 3, Uppsala Fax +46 (0)18 471 4687	

Deviating time-to-onset in predictive models

- detecting new adverse effects from medicines

Caroline Wärn

Populärvetenskaplig sammanfattning

Effektiva läkemedel kommer sällan utan biverkningar. Under de senaste decennierna har antalet läkemedel och omfattningen av deras användning ökat, något som har förbättrat livskvaliteten och förlängt livslängden hos befolkningen. I samma takt har biverkningar blivit vanligare och är idag en av de tio främsta orsakerna till ohälsa, sjukdom och dödsfall i de industrialiserade länderna. För att kunna utföra en korrekt avvägning mellan fördelar och nackdelar med en läkemedelsbehandling innan läkemedlet sätts in så måste risken för eventuella biverkningar vara känd. Detta har föranlett utvecklingen av metoder som genom ett stort antal biverkningsrapporter identifierar potentiella biverkningar så tidigt som möjligt.

Det här examensarbetet syftar till att utvärdera en ny metod för identifiering av läkemedelsbiverkningar hos icke-vacciner då metoden tidigare har givit lovande resultat för vacciner. Principen är att identifiera avvikelser i tidsperioden mellan det tillfälle då ett läkemedel administreras till dess att en biverkning uppstår. Metodens prediktiva förmåga testades separat, men även tillsammans med den multivariata prediktiva modellen vigiRank. Prestandan hos den nya metoden visade sig inte vara tillräckligt hög för att inkluderas i vigiRank och gav inte heller tillräcklig prediktiv förmåga för att användas som ett fristående verktyg. Därmed anses den vara av begränsat värde för prediktering av biverkningar från icke-vacciner.

Examensarbete 30 hp
Civilingenjörsprogrammet Molekylär bioteknik,
inriktning Bioinformatik

Uppsala universitet, juni 2015

CONTENTS

1	INTRODUCTION.....	9
1.1	Background.....	9
1.1.1	<i>Pharmacovigilance.....</i>	9
1.1.2	<i>The Uppsala Monitoring Centre.....</i>	10
1.1.3	<i>Disproportionality analysis.....</i>	11
1.1.4	<i>vigiRank.....</i>	13
1.1.5	<i>Deviations in time-to-onset.....</i>	14
1.2	Aims and objectives.....	15
2	METHODS AND IMPLEMENTATION.....	16
2.1	Data extraction.....	16
2.1.1	<i>Reference set with emerging signals.....</i>	16
2.1.2	<i>Additional data filtering for emerging signals.....</i>	17
2.1.3	<i>Established signals.....</i>	18
2.1.4	<i>Implementation of data extraction and filtering.....</i>	18
2.2	The Kolmogorov-Smirnov test.....	19
2.2.1	<i>Procedure.....</i>	19
2.2.2	<i>Implementation of the Kolmogorov-Smirnov test.....</i>	21
2.3	vigiRank.....	22
2.4	ROC curve analysis.....	22
2.4.1	<i>Implementation and comparisons.....</i>	24
2.5	Development environments.....	24
3	RESULTS AND ANALYSIS.....	25
3.1	Handling of the first day of the month.....	25
3.2	Length of TTO time window.....	26
3.3	P-value approximation.....	26
3.4	vigiRank.....	28
3.5	Prediction of established signals.....	28
4	DISCUSSION AND CONCLUSIONS.....	30
5	ACKNOWLEDGEMENTS.....	33

6	REFERENCES	34
7	APPENDICES	37
	Appendix A.....	38
	Appendix B.....	43
	Appendix C.....	50
	Appendix D.....	54

LIST OF ABBREVIATIONS AND ACRONYMS

ADR	Adverse Drug Reaction
AUC	Area Under the Curve
CAP	Centrally Authorized Product
CDF	Cumulative Distribution Function
IC	Information Component
ICSR	Individual Case Safety Report
KS	Kolmogorov-Smirnov
OE	Observed to Expected
ROC	Receiver Operating Characteristic
SPC	Summary of Product Characteristics
TTO	Time-To-Onset
UMC	The Uppsala Monitoring Centre
WHO	The World Health Organization

1 INTRODUCTION

Today's extensive use of medicines also brings with it the risk of adverse drug reactions. In the developed world, adverse drug reactions are ranked as one of the top 10 causes of illness and death [1], costing almost \$180 billion annually only in the United States [2]. Approximately 80% of the American adults use at least one medication and 25% use five or more [3]. Both the usage and the number of medications increases with age. This is of great concern since the human population is growing older each and every year. Furthermore, the prescription rate has increased to the point where 64% of all visits to a doctor result in prescriptions [4]. However, the development of new drugs has led to a healthier and longer living population, showing the need to decrease the risk of adverse drug reactions when possible and increase the capability of performing an accurate benefit-to-harm assessment.

One major step in this process is detecting new adverse drug reactions as early as possible to limit the harm caused by them. To do this, multiple tools and methods have been developed for analysis of individual case safety reports. The most widely used is disproportionality analysis, but lately, other alternatives have been shown to complement [5] or even outperform [6] this type of analysis.

A novel approach to signal detection for vaccines was presented by Van Holle et al. [5]. The concept was to identify vaccine-event pairs that deviated in their reported time from vaccine administration to event onset compared to other vaccines or events. This approach has shown promising results for vaccines, but has not yet been tested for non-vaccine drugs. Another approach has been proposed by Caster et al. [6] where a predictive model including five different variables was used in the signal detection process for drugs. If the two approaches complement each other it could possibly lead to synergistic effects, improving the overall efficacy of signal detection. However, to test whether this is the case, the predictive power of deviating time-to-onsets must first be evaluated for non-vaccines. This is the objective of this project.

1.1 Background

1.1.1 Pharmacovigilance

Effective medical therapies usually come with additional unintended effects. These are the major concerns of pharmacovigilance. Pharmacovigilance is defined by the WHO as “the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other possible drug-related problems” [7]. In addition, pharmacovigilance has also come to include herbals,

traditional and complementary medicines, vaccines, biologicals, blood products and medical devices.

Drugs are monitored during both premarketing development and postmarketing, making sure that possible adverse drug reactions (ADRs) can be detected throughout the duration of their use [8]. It is widely known that the main focus of clinical trials is on the efficacy of the drug rather than on the safety profile [9]. Moreover, most clinical trials only test the drug during a short period of time on a small group of homogenous patients, excluding those with the greatest risk of experiencing problems (e.g. children and elderly) [10]. The small number of patients makes it impossible to detect serious ADRs occurring at low frequencies. If an adverse reaction for a drug occurs with a 1:20 000 frequency, the drug would need to be exposed to approximately 60 000 patients to be 95% certain that the ADR would occur at least once. Most clinical trials do not include more than 5 000 patients, while the market may consist of several millions. Rare ADRs will therefore only become known after the drug has entered the market, making postmarketing surveillance a necessity.

Postmarketing surveillance has its own limitations because of missing data and underreporting. The underreporting occurs in two steps, firstly, the patients fail to report adverse reactions to their doctors, and secondly, doctors fail to report adverse reactions to national authorities. The median underreporting rate of ADRs has been shown to be as high as 94% [11]. However, more severe ADRs are normally reported more frequently by general practitioners, resulting in a median underreporting rate of 80%.

Lack of data gives both the health practitioner and the patient a false sense of security about the drugs that are prescribed and used. The main concern in the science of pharmacovigilance is to, as early as possible, detect previously unknown ADRs. However, it is also used to detect inappropriate prescription and administration as well as getting further insight into the pharmacological mechanisms that cause the adverse reactions [8]. These activities lead to a better understanding and assessment of the benefit-to-harm balance, helping doctors and patients to select the most appropriate medicine as well as helping regulatory agencies to decide whether the medicine should be discontinued or not.

1.1.2 The Uppsala Monitoring Centre

The World Health Organization (WHO) recognized the need for drug safety monitoring in 1962, six months after the thalidomide disaster [12]. Thalidomide was initially prescribed as a sedative but later promoted for use by pregnant women as a treatment for morning sickness and nausea. Four years after its introduction on the European market, thousands of infants had been born with abnormalities such as phocomelia and micromelia (malformed and shortened limbs) and the drug was withdrawn from sale.

The thalidomide incident resulted in multiple actions taken by both physicians and health authorities, and the WHO was requested to take on the leading role for

international drug safety surveillance [12]. A pilot project started in Alexandria, Virginia, USA, in 1968 with the purpose of developing an international system for detection of previously unknown adverse effects of medicines. Three years later it moved to the WHO headquarter in Geneva, Switzerland, to what came to be known as the “WHO Research Centre for International Monitoring for Adverse Reactions to Drugs”. The project was initially financed by the Government of the USA, but in 1978 the financial responsibility was taken over by the Swedish government. This led to the establishment of the centre now called the Uppsala Monitoring Centre (UMC), a WHO collaborating centre for drug monitoring. However, the financial support ended in 2001 when the UMC was able to generate its own income by selling a drug dictionary among other products and services.

Today, the UMC is the WHO collaborating centre for international drug monitoring with the main aim to identify pharmacovigilance signals as early as possible. The UMC develops and maintains the world’s largest database, VigiBase®, containing more than 11 million individual case safety reports (ICSRs) from 122 countries as of June 2015. It was developed in the mid-1990s and is updated continuously [13]. The ICSRs are received from national centers around the world, which in turn are provided with reports from pharmaceutical companies, health professionals and, in some countries, patients [14]. When an ADR is found by the UMC, it is communicated to the pharmaceutical company responsible for the medical product and to national centers. The regulators in each country then have the power to take firm action, e.g. adding the adverse reaction to the label of the medical product or issuing a warning.

1.1.3 Disproportionality analysis

One approach for detecting new ADRs is by looking at variations in the rates of events. Some events might be more frequently reported than what would be expected overall or in specific time periods, regions, age groups etc. Disproportionality analysis is one tool to measure the deviations of reporting frequencies of a drug-event combination from the baseline behavior. This type of analysis overcomes one of the major limitations of ICSR data, that is, the lack of reliable estimates of the exposed population. It has therefore long been the accepted standard for highlighting possible ADRs for in-depth clinical assessment [15].

Disproportionate reporting can be measured in various ways. One approach is by comparing the observed number of reports for a specific drug-event combination with the expected number of reports for that combination, something that is called the observed to expected (OE) ratio [16]. When the differences between the two groups are large, the association between the drug and the event tends to be of importance. The algorithm for disproportionality analysis by the OE ratio is outlined below.

Let the following contingency table show the frequencies of reports containing a specific drug (x) and/or a specific event (y).

	y	not y
x	a	b
not x	c	d

The OE ratio for the pairwise association can then be calculated as the ratio of $f(y | x)$, i.e. the relative frequency of the event on reports where the drug is present, to $f(y)$, the marginal relative frequency of the event. This can be formulated as follows:

$$\frac{f(y | x)}{f(y)} = \frac{a / (a + b)}{(a + c) / (a + b + c + d)} \quad (1)$$

By reordering the factors, the expression below is obtained. Here, the observed number of events is given by the numerator and the expected number of events is given by the denominator.

$$\frac{O}{E} = \frac{a}{(a + b) (a + c) / (a + b + c + d)} \quad (2)$$

The above definition of disproportionate reporting has one important limitation, the marginal frequency of the event is the determining factor for which values the OE ratio can take on. The OE ratio cannot exceed $1 / f(y)$, which implies that if the event is reported with a high marginal frequency, the obtained OE ratio would always be low. This definition is therefore only useful when the event is rare, which is usually the case in Vigibase. However, when the expected number of events is low, the OE ratio becomes very volatile, changing values drastically even for very small alterations in the expected number of events. This is a major drawback for drug safety data, since rare events may be of great importance.

To handle this behavior, the OE ratio can be subjected to statistical shrinkage. This is done purely for the purpose of reducing the effects of sampling variation and stabilize the OE ratio when the expected number of events is low. A simple form of shrinkage transformation, as proposed by Norén et al. [16], is outlined below. If the observed number of events is denoted O and the expected number of events is denoted E , then the expression for the shrinkage transformation can be formulated as follows, when conditioned on E :

$$\frac{O + \alpha_1}{E + \alpha_2} \quad (3)$$

In practice, the impact of the shrinkage would be equal to adding α_1 observed events and α_2 expected events, driving the OE ratio towards α_1 / α_2 . When $\alpha_1 = \alpha_2$, the OE ratio is biased towards one, indicating proportionate reporting, i.e. no differences in the observed and expected number of reported events. The risk of highlighting inaccurate associations is thereby decreased. If prior knowledge of the specific association exists, the α_1 / α_2 ratio may be adjusted accordingly to provide shrinkage in one direction or the other. The impact of the shrinkage is determined by the number

of observed and expected events. When the number of events is large, the impact of the shrinkage will be small, leading to reduction in variation only when needed. Based on the data in VigiBase, the values have been set to $\alpha_1 = \alpha_2 = 0.5$ for the implementation currently in use at the UMC to bias the ratio towards the baseline value one, while limiting the impact of the shrinkage. The base 2 logarithm is then applied to the OE ratio with shrinkage to obtain the final measure, giving both direction and strength of the association. This measure is called the Information Component (IC) [6].

$$IC = \log_2 \frac{O + 0.5}{E + 0.5} \quad (4)$$

1.1.4 *vigiRank*

Even though disproportionality analysis is the most common method for finding new ADRs from ICSRs, it is limited to what it measures. It is solely based on the number of reports and does not take report quality and content into consideration. This was the main reason for why the UMC decided to develop *vigiRank* [6], a predictive tool based on lasso logistic regression which is currently used as the standard at the UMC today.

vigiRank was developed using 13 different variables to capture different aspects of the reports and reporting patterns. These variables were selected by Caster et al. [6] through consensus with pharmacovigilance experts. The proposed variables were *Disproportional reporting*, *Informative reports*, *Narrative*, *Dechallenge*, *Rechallenge*, *Causality assessment* (including two separate variables), *Solely reported*, *Multiple reporting elements*, *Recent reporting*, *Geographic spread* and *Time trend*. The final variable going into the lasso logistic regression was *Time-to-onset*. This variable measured the number of reports within a reasonable time span (in this case, 90 days) between drug administration and ADR onset. However, the implementation of the time-to-onset variable was only a crude attempt to capture this type of information, as noted in the article.

The 13 variables went into a lasso logistic regression model, fitted based on 5 544 samples of which 264 were positive controls, i.e. emerging safety signals. Variables which were not binary, underwent a transformation while binary variables remained as they were. A logistic regression model is generally formulated as:

$$\log \frac{P(y | x)}{1 - P(y | x)} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad (5)$$

In this formula, β_i can be seen as the log odds ratio of the predictor variable x_i while p denotes the number of variables. It can be reformulated by solving for $P(y | x)$:

$$P(y | x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)} \quad (6)$$

The logistic regression was regularized by constraining the coefficients using an L_1 constraint and creating a lasso logistic regression. By using the lasso method, excessive predictor variables are penalized and might be discarded from the equation, resulting in a less complex model [17]. Elimination of predictor variables occur when the variable coefficient is estimated to zero. The constraint imposed by using the lasso method can be formulated as follows:

$$\sum_{i=1}^p |\beta_i| \leq t \quad (7)$$

Here, the amount of shrinkage and the size of the model is determined by the tuning parameter t . This parameter was set so that the chosen model was the largest possible model without negative coefficients since these were assumed to suggest overfitting.

The model used by the UMC is based on five variables including *Informative reports*, i.e. a measure of the completeness of a report [18] and *Narrative*, indicating whether the report has any free text information. It also includes *Disproportional reporting* (as described in 1.1.3 Disproportionality analysis, with extensions to local patterns [19]), *Recent reporting* and *Geographic spread*. Estimations of model performance indicated that vigiRank can outperform disproportionality analysis alone [6].

1.1.5 Deviations in time-to-onset

Another approach for signal detection, specifically for vaccines, has been presented by Van Holle et al. [5]. They recognized that some events usually occur within a specific time window post-vaccination. They concluded that the distribution of time-to-onsets might differ for a specific combination compared to others. In their study, they used the two-sample Kolmogorov-Smirnov (KS) test (see 3.2 The Kolmogorov-Smirnov test) to analyze time-to-onset (TTO) data for vaccine-event pairs. TTO was defined as the number of days from vaccine administration to event onset. The purpose of using the KS test was to compare the shapes of two distributions of data. If they were significantly different, the adverse event was defined as a statistical signal. The underlying assumption was that most vaccine-event pairs in safety report databases are not causally related and because of this, one of the two distributions should be dominated by reporting biases and noise, defining the overall shape of the distribution.

Van Holle et al. [5] implemented two versions of the test; “between events” and “between vaccines”. When performing the “between events” test, a specific vaccine-event pair was compared to combinations including the same vaccine with all other reported events. To the contrary, when performing the “between vaccines” test, the specific vaccine-event pair was compared to combinations including the same event, but reported after administration of other vaccines. Eight different vaccines were

used, which corresponded to a diverse set of indications. Taken together, these made up more than half of the reports in their database. The events for the positive controls were taken from the summary of product characteristics for each vaccine, i.e. the associations between the vaccines and their respective events were established.

The studies performed by Van Holle et al. [5, 20-21] involved careful evaluations of the impact of different TTO windows (30, 60 and 90 days) and different significance levels (0.01, 0.05, 0.10, 0.20, 0.50 and 0.99). The two-sample KS test was used as implemented in the statistical software SAS[®] after appropriate filtering of the data. Their results were highly promising, indicating that the test could be used as a complement to the standard disproportionality analysis.

Van Holle et al. [5] presents a new way of using TTO data in large datasets. However, the studies were performed on vaccine data only, raising the question whether the results would be transferable to non-vaccines. If this method would work for non-vaccines as well, it could be a good replacement for the crude estimate of TTO patterns used in the original implementation of *vigiRank* and possibly increase its performance.

1.2 Aims and objectives

The aim of this Master's thesis project is to test the performance of a predictor based on deviating time-to-onset patterns for predictions of adverse drug reactions from non-vaccines. The performance is compared with that of disproportionality analysis and *vigiRank*. Furthermore, the predictor is added to *vigiRank* to investigate whether it gives sufficiently independent predictive power to be included as a component of the model.

2 METHODS AND IMPLEMENTATION

2.1 Data extraction

Two datasets were used for evaluation of the KS test. The first dataset came from Alvarez et al. [22] and contained emerging safety signals while the other contained established safety signals. The first dataset was initially filtered to correspond to the dataset used for the original fitting of *vigiRank*. However, some additional filtering had to be done, resulting in a much lower number of combinations used for the evaluation of the KS test compared to those used for the original fitting of *vigiRank*. Identical filtering was applied to both datasets for comparative purposes.

2.1.1 *Reference set with emerging signals*

To evaluate the predictive power of the KS test, the outcome of the ingoing data must be known and assumed to be correct. Furthermore, the controls must include both combinations with a confirmed association (i.e. positive controls) and combinations where no such association has been seen (i.e. negative controls). By doing this, one can obtain objective indications to whether the predictor is expected to be effective on similar data or not.

The importance and relevance of distinguishing emerging safety signals from established casual associations has previously been thoroughly investigated [23]. When an ADR for a drug has become established, the number of reports for that association tends to increase. In addition, the TTO might be reported as the expected TTO in cases when the reporter is uncertain of the actual time period, giving a bias towards the expected TTO rather than the actual for that combination.

To circumvent this issue, emerging safety signals were used as controls. The controls were taken from a dataset compiled by Alvarez et al. [22]. It contained 532 historical safety signals from September 2003 to March 2007 for 267 centrally authorized products (CAPs) from the European Medicine Agency. Combinations including vaccines were excluded from the controls due to their very different nature. Furthermore, only combinations reported from at least two different countries were considered in the reference set.

From the original 532 safety signals, 264 signals from 65 CAPs were left as positive controls. The negative controls were randomly selected from the original 267 CAPs where the event was not included in the 2012 European summary of product

characteristics (SPC) for that drug. The number of negative controls was selected at a 20:1 ratio to the positive controls, resulting in 5 280 combinations.

Reports containing either the drug or the ADR from the 5 544 selected controls were extracted from a version of Vigibase backdated to 31 December 2004, when most of the signals were still emerging. Reports where the drug was listed as concomitant were excluded while reports containing drugs listed as suspected or interacting were kept. Furthermore, suspected duplicated reports were also removed. Duplicates may be present due to multiple reporting of the same event by different sources or due to erroneous linking between events and follow-up reports of these events [14]. The dataset thereby included the exact same data as was used for the original fitting of vigiRank [6].

2.1.2 Additional data filtering for emerging signals

To prepare the dataset for the KS test, some additional filtering had to be done. First, all reports missing either one of the two dates necessary to compute a TTO value were defined as missing data and were not used in the KS test. This meant excluding 59% of the dataset. In addition, reports containing incomplete dates were also excluded since the TTO values were given at a precision of days. TTO values were then calculated for each reported combination. The TTO was defined as the number of days between drug administration and event onset including these two dates. Three different variations of the dataset were created using different time windows. These variations included reports containing TTO values within the intervals 0-30, 0-60 and 0-90 days respectively. The intervals were chosen to correspond to the study made by Van Holle et al. [21] for easy comparison.

When multiple prescriptions of the same drug were reported together, the one with the shortest TTO was selected for the KS test since this was how the UMC had handled the situation before. These multiple prescriptions may e.g. occur when a drug is administered as a bolus intravenously to quickly raise the concentration of the drug in the blood while then afterwards continuing with oral administration at a lower dosage. The products are different but mapped to the same substances in Vigibase, causing multiple prescriptions for identical instances of combinations.

Some national center systems autocomplete incomplete dates. When only the month is given, some systems automatically fill in the 1st or the 15th of that month. When only the year is given, the systems might autocomplete the dates to the first of January or to the middle of June. The majority of reports being autocompleted to the first day of the month came from the USA who reported ten times more instances on the first day compared to every other day in a month. Similar patterns, although not as prominent, could be seen for France, who had submitted three times more reports containing dates including the 15th of a month compared to other days. This issue necessitated a structured way of handling the data.

When calculating TTO, the UMC has earlier changed all first days in a month into the 15th of that month, limiting the deviating number of days to 15 instead of 31.

However, this was not an obvious solution for this application since a greater precision was desired. To find the best solution for this application, three different variations of the dataset were created. The first variation contained all original data, including dates as they were entered into the system. This meant that most of the dates including the first day in a month would be wrong and possibly disturbing further analysis. In the second variation, all first days in each month were excluded to avoid incorrect dates. However, this meant that some correct dates were also excluded, making the number of reports used for the KS test smaller than necessary. The last variation handled the first days as the UMC had previously handled them, including the reports but adjusting the dates to the middle of the month. A total of nine variations of the dataset were created, containing all combinations of time windows and date handlings.

2.1.3 Established signals

For comparative purposes, a reference set based on established safety signals was also generated. This dataset had previously been used by the UMC [23] and contained a total of 31 414 combinations. 16 091 of these were defined as positive controls, where the events were taken from the SPC for CAPs in Europe. The other 15 323 combinations were defined as negative controls, including ADR terms not listed in the drug's SPC in Europe as of 2012. Exclusion of combinations for the KS test was performed in the same manner as described above, i.e. only non-vaccines and combinations being reported from at least two different countries were considered.

The data filtering procedure was similar to the one performed for emerging signals. Reports containing drugs listed as concomitants were excluded as well as suspected duplicated reports. 53% of the reports were defined as missing data since they lacked at least one of the two dates necessary for calculation of the TTO value. Furthermore, incomplete days were also excluded. Only reports with dates not containing the first day of a month and only events occurring within the first 60 days were included. Finally, multiple prescriptions were handled by selecting the one with the shortest TTO value as described above.

2.1.4 Implementation of data extraction and filtering

The data extraction and filtering was implemented as a stored procedure in SQL (Appendix A). It was designed to extract data from Vigibase and prepare it for the KS test as previously described. The time window determining which reports to include was given as a user-defined input parameter as well as how to handle the first day in each month. The output table contained all entries in Vigibase encompassing combinations including either the drug or the ADR from the reference set and their calculated TTO values.

2.2 The Kolmogorov-Smirnov test

The two-sample Kolmogorov-Smirnov test is a nonparametric hypothesis test developed in the 1930's by Kolmogorov [24] and Smirnov [25]. One of its biggest advantages over similar tests is that it makes no assumptions about the underlying distribution of data. This makes it an appropriate choice when the distributions are not known and/or the data cannot be guaranteed to be normally distributed (Fig. 1a). The test is generally used for detection of differences in distributions, taking medians, variances, shifts, kurtosis and overall shapes into account [26]. However, it is well known that the test is more sensitive to these deviations in the central parts of the distributions rather than at the tails.

Two conditions must be fulfilled for the test result to be valid; the data must be drawn from a continuous population and the two samples must be based on independent variables [27]. In such cases, the two-sample KS test can be used to determine whether the data in two samples are drawn from the same probability distribution (null hypothesis, H_0) or from two different distributions (H_1). This can be written as below where $F(x)$ is the distribution function describing population 1 of size n , containing the random observations (X_1, X_2, \dots, X_n) and where $G(x)$ is the distribution function describing population 2, of size m , containing the random observations (Y_1, Y_2, \dots, Y_m) .

$$\begin{cases} H_0: F(x) = G(x) & \text{for each } x \\ H_1: F(x) \neq G(x) & \text{for at least one } x \end{cases}$$

2.2.1 Procedure

The first step in the KS test is to compute the empirical cumulative distribution functions (CDFs) of the two samples, here denoted $S(t)$. Given n observations of the random variable X , ordered in ascending order as $X_1 \leq X_2 \leq \dots \leq X_n$, the empirical CDF is defined as the fraction of observed X 's to the left of the number t , for every real number t . This can be formulated as follows:

$$S(t) = \frac{1}{n}(\text{number of obs. } X' s \leq t) = \frac{1}{n} \sum_{i=1}^n I(x_i \leq t) \quad (8)$$

The empirical CDF is a step function which is constant between consecutive x values and increases at each data point with the fraction $1 / n$. Every CDF starts at zero and ends at one, but varies in between these endpoints. This variation distinguishes one cumulative distribution function from another and can be measured in multiple ways. The two-sample KS test measures the divergence between two empirical CDFs as the supremum value of the absolute difference between each data point (Fig. 1b), defined as the KS test statistic D [28].

$$D = \sup_{-\infty < x < \infty} |S_1(x) - S_2(x)| \quad (9)$$

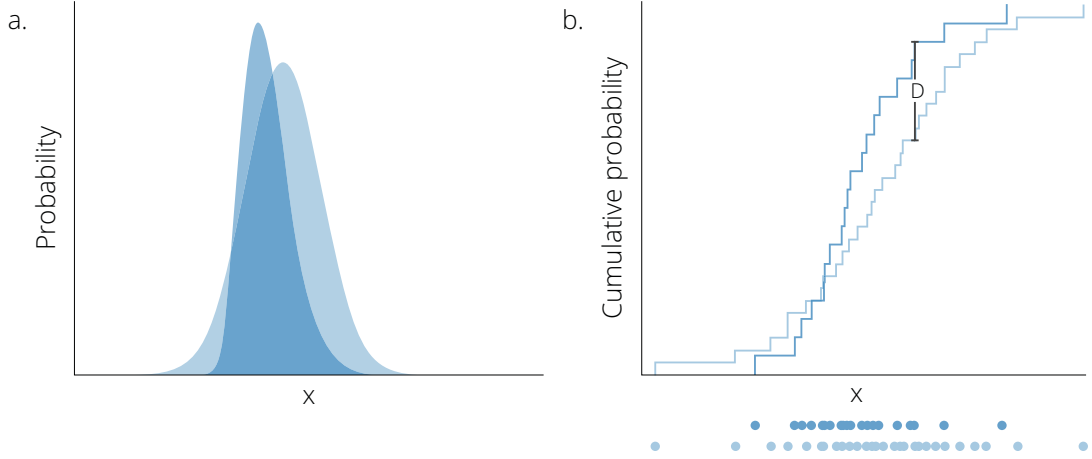


Figure 1. Probability distributions. **a.** Two different probability distributions comprising a gamma distribution (blue) and a normal distribution (light blue). **b.** Two empirical cumulative distribution functions generated by 20 random samples from the given gamma distribution (blue) and 30 random samples from the given normal distribution (light blue). The sample points are plotted along the x axis below the graph. The D-value is defined as the supremum vertical absolute distance between the two curves.

One approach for evaluating the null hypothesis is to compare the obtained D-value with the critical d-value, i.e. the D-value of the test when the null hypothesis cannot be rejected [29]. The critical d-value can be approximated with the Smirnov approximation [25], (formula 10-11). If the obtained D-value is greater than the critical d-value, the null hypothesis can be rejected.

The probability that the computed D-value is greater than the critical d-value under the null hypothesis is defined as the p-value of the test. The exact p-value includes complex and involved calculations and is therefore often replaced with an asymptotic value which becomes more accurate as samples grow larger. There have been a multitude of suggestions of how to best approximate the p-value, but no consensus has been reached. Simard and L'Ecuyer [30] make a thorough investigation of possible approximations, discussing their respective strengths and weaknesses. They conclude that multiple approximations should be used in various regions of the (n, x) space. Furthermore, they note that all of the tested approximations perform very poorly on small sample sizes, where tabularized exact values may be used instead. What follows is a description of one implementation of the p-value approximation. Two other implementations can be viewed in Appendix D. The approximation is based on the effective number of values, N , where N_1 is the number of values in the first sample and N_2 is the number of values in the second sample.

$$N = \frac{N_1 N_2}{N_1 + N_2} \quad (10)$$

The following approximation is based on formulas presented by Smirnov [25] and is implemented by the statistical software R [31] as the function `ks.test()` (package `dgof`)

[32] as well as by SAS[®] as the procedure `npar1way()` [33]. In R, the approximation is used when specified or when the product of the two sample sizes is greater than 10 000. Otherwise, an exact method is provided. Although giving the option of using an exact method for calculation of the p-value, this implementation has been criticized [30]. The approximation has been shown to be inaccurate even for fairly large samples sizes, while the exact method is too slow to be useful. When performing the calculations, the KS test statistic is first standardized to obtain the critical d-value under the null hypothesis (λ) which converges to the Kolmogorov distribution for large sample sizes. This asymptotic distribution was originally derived by Smirnov in 1939 [25]. The value is then used in the approximation of the p-value. The complement of the cumulative distribution function of the KS distribution enters into the approximation of the p-value and is defined by the series in equation (12).

$$\lambda = D\sqrt{N} \quad (11)$$

$$P(D > d) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2\lambda^2} \quad (12)$$

The p-value is used to determine if the difference between the distributions is significant enough to reject the null hypothesis, based on a specified significance level α . If the p-value is less than α , the null hypothesis can be rejected and vice versa.

The calculated p-values tend to be conservative, i.e. the p-values are strongly biased upward. Most approximations estimate the parameters from the data rather than from the sample, something that is known to cause conservative p-values [34]. Furthermore, using discrete values instead of continuous as well as performing the test on tied data contributes to a possibly misleading result. As earlier mentioned, the approximations do not perform well on small sample sizes, adding to the issue. This means that even if an exact method is used, the p-value might still be erroneous due to the composition of the underlying data.

Another important notice to make is about the relative sizes of the two samples being compared. It has been shown that if one sample is very large while the other is relatively small, the performance will not be improved by adding more observations to the larger group [35]. Paradoxically, the opposite is true. The additional information contained in a larger sample will not give any benefits, instead the added information will decrease the power of the KS test. It has also been suggested to use tables with exact values when the samples are strongly imbalanced since the approximations will be poor for this scenario [29]. However, no tables do yet exist for the extreme imbalance present in *VigiBase*.

2.2.2 Implementation of the Kolmogorov-Smirnov test

The two-sample KS test was implemented as a stored procedure in SQL (Appendix B). The possibility to select which type of test to perform (i.e. between drugs or between events) was implemented as an input parameter, as well as the possibility to

specify a significance level. Which equations to use for approximation of the p-value could easily be changed in the code (see 2.2.1 Procedure). The output table contained the drug-ADR pair together with their calculated KS statistic D, the approximated p-value and a hypothesis based on the given significance level. Furthermore, cautionary notes were given for combinations that did not fulfill predefined criteria regarding the number of reports in each sample. This criteria made sure that the two sample groups contained a minimum number of reports. The output from the KS test procedure was then used in another stored procedure (Appendix C) to obtain performance measures, i.e. sensitivity and specificity (see 2.4 ROC curve analysis).

2.3 vigiRank

The KS test result was also integrated into the fitting process of vigiRank. This was done by retrieving the data for all 13 original variables, excluding TTO 90 days (see 1.1.4 vigiRank). These data were based on the complete dataset of emerging signals without additional filtering (see 2.1.2 Additional data filtering for emerging signals). The variable TTO 90 days was replaced with the test result from the KS test. This result was based on reports where the ADR onset occurred within the first 60 days and where all reports containing the first day of a month were excluded. All other reports were classified as missing data. This variation of the dataset contained 3 826 combinations. However, when setting the criteria of only including combinations where the KS test result had been based on two samples containing at least two reports each, only 2 416 combinations were left including 162 positive controls and 2 254 negative controls. For comparative purposes, the complete dataset of 5 544 combinations was also used for refitting of vigiRank by setting the test result to 0 for all combinations with missing values.

The results up to this point had indicated a higher predictive power using the D-value of the KS test compared to using the p-value. In addition, uncertainties in the calculation of the critical d-value and in the p-value approximation existed as previously discussed. Because of this, the raw D-values were used as the test result in the refitting process.

To investigate whether the KS test measured the same feature as some other variable in vigiRank, a correlation test was performed between the D-value and all original 13 variables. It was implemented using the R `cor()` function and the Pearson product-moment correlation coefficient. The correlation coefficient can take on values between and including -1 and 1. When two variables are independent of each other, the correlation coefficient is 0, when there is a total positive correlation or a total negative correlation, the coefficient will be 1 and -1 respectively.

2.4 ROC curve analysis

When comparing a model's prediction with the true outcome, measures of sensitivity and specificity can be obtained. These are calculated from the number of combinations where the events are correctly identified as ADRs (true positives, TP),

where they are incorrectly identified as ADRs (false positives, FP), where they are correctly rejected as ADRs (true negatives, TN) and where the events are incorrectly rejected as ADRs (false negatives, FN). Sensitivity is the true positive rate, i.e. the fraction of ADRs which are identified as such. To the contrary, the specificity is defined as the true negative rate, i.e. the fraction of events which are not ADRs and rejected by the predictor.

$$\text{sensitivity} = \frac{TP}{TP + FN} \quad (13)$$

$$\text{specificity} = \frac{TN}{TN + FP} \quad (14)$$

The measures are dependent on the threshold value determining predictor outcome, e.g. the p-value of a statistical test (Fig. 2a). For a more thorough evaluation of the performance, the sensitivity and specificity can be calculated for every possible value of the threshold. This will provide a foundation for comparing different predictors with each other as well as to support the selection of an appropriate threshold [36]. To do this, sensitivity is plotted against 1-specificity, resulting in a so called receiver operating characteristic (ROC) curve (Fig. 2b). The diagonal of this graph (where sensitivity + specificity = 1) corresponds to a completely uninformative test, i.e. a predictive power equal to random chance. To the contrary, a test with perfect discrimination will result in a ROC curve passing through the upper left corner.

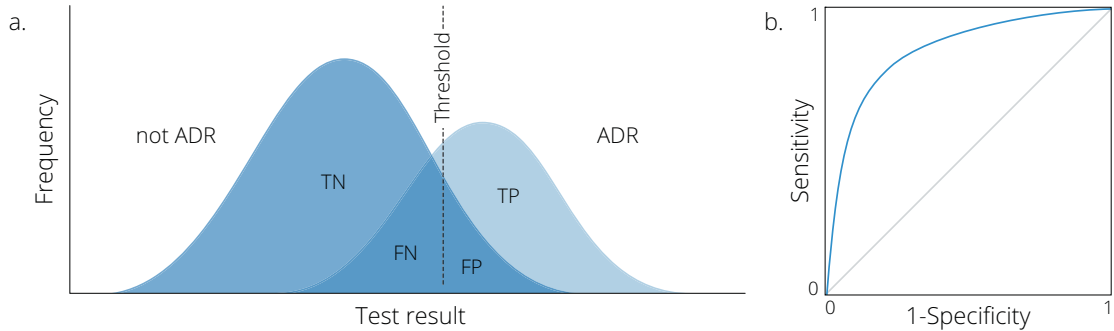


Figure 2. Concept of ROC curves. **a.** The two distributions describe the positive and negative combinations. The threshold determines at which point the combinations should be separated leading to different proportions of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). **b.** A receiver operating characteristic curve describing the tradeoff between sensitivity and specificity.

Another way of describing the performance is in terms of the area under the ROC curve (AUC). The AUC value corresponds to the probability that a randomly chosen positive combination has a higher probability of being positive than a randomly chosen negative combination. A random predictor would have an AUC value of 0.5, meaning that a random positive combination and a random negative combination would have the same probability of being classified as positive, while a perfect predictor would have an AUC of 1.0, i.e. only the positive combinations will be predicted as such.

2.4.1 Implementation and comparisons

ROC curve analysis was performed for the KS test results (both D-value and p-value) of all nine variations of the dataset containing emerging signals. The variations contained all combinations of the time windows (30, 60 and 90 days) and the handling of the first days in each month (keeping, excluding and changing). ROC curves were also generated for the dataset with established signals for the time window 60 days where the reports containing the first day in a month were excluded. Finally, ROC curves were generated for the refitted *vigiRank* (described in 2.3 *vigiRank*) as well as for the $IC_{0.25}$ (the lower limit of the 95% credibility interval of the information component) and the raw number of reports for each combination. The last two methods were based on all reports without the additional filtering. All ROC curves were compared visually and by their AUC values.

2.5 Development environments

The two-sample KS test was implemented in Microsoft SQL server 2014. Since the UMC uses the 2012 version, it was made to be backward compatible to that version but not any further. The SQL server is a relational database management system developed by Microsoft using Transact-SQL as the main query language. This system is used to store and manage the databases at the UMC, including the ICSR database. Due to convenience, the test was implemented in the same environment to avoid unnecessary transfers of large datasets.

The refitting of *vigiRank* was performed in R (version 3.1.2) since this was the environment of the original *vigiRank* implementation. R is an open source software used for statistical computing and plotting and is currently developed by the *R Development Core Team*. One of the main strengths of R is the user-created packages which rapidly extends the capabilities of the software. However, in addition to R, both SAS (University Edition) and MATLAB (R2015a, version 8.5) were used for validation purposes, making sure that the implementations of their respective approaches for approximating the p-values were correct. MATLAB was also used for the ROC curve analyses and plotting.

3 RESULTS AND ANALYSIS

3.1 Handling of the first day of the month

Receiver operating characteristic curves were generated for the three variations of first day handlings for the dataset containing emerging signals (Fig. 3). These were all based on the KS test statistic D as the determining variable. The curves correspond to variations where all the first days of a month were kept as they were entered into the system (including 2 770 combinations with 179 positive controls and 2 591 negative controls), where they were excluded (2 416 combinations, 162 positive and 2 254 negative controls) and where they were changed to the 15th of the same month (2 796 combinations, 180 positive and 2 616 negative controls). The time window for all three variations was set to 60 days.

The performances of the predictors for the different variations were very similar, however, the AUC value was slightly better for the variation where all first days were excluded (AUC = 0.624) compared to the variations where the first days were kept (AUC = 0.617) and where they were changed (AUC = 0.613). This might indicate that the removal of incorrect data increases performance more than a smaller dataset decreases it. Furthermore, excluding potentially incorrect dates makes more logical sense since the KS test should be sensitive to small variations in the distributions of TTO values.

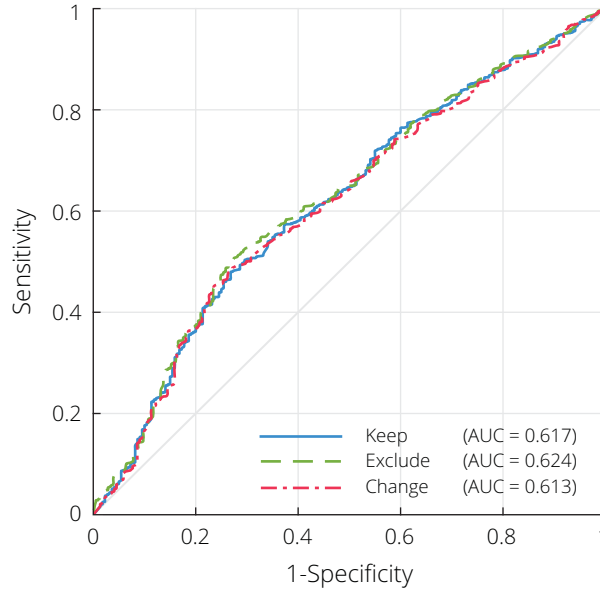


Figure 3. Handlings of the first day in a month. Receiver operating characteristic curves for three different handlings of autocompleted dates. These include keeping all the first days in each month as they were entered, excluding all the first days in each month and changing the first days to the middle of the month.

3.2 Length of TTO time window

Three different time windows for the emerging signals were selected to investigate their impact on performance (Fig. 4). The curves were based on the KS test statistic D for combinations represented by reports where the first day of each month was not present. The curve for the 30 days time window was based on 2 160 combinations, where 142 were positive controls and 2 018 were negative controls. The curve for the 60 days time window was based on 2 416 combinations whereof 162 were positive controls and 2 254 were negative controls while the result for the 90 days time window was based on 2 563 combinations, where 170 were positive controls and 2 393 were negative controls. The curves were very similar with a slightly higher AUC value for the 60 days dataset (AUC = 0.624) compared to 30 days (AUC = 0.619) and 90 days (AUC = 0.615).

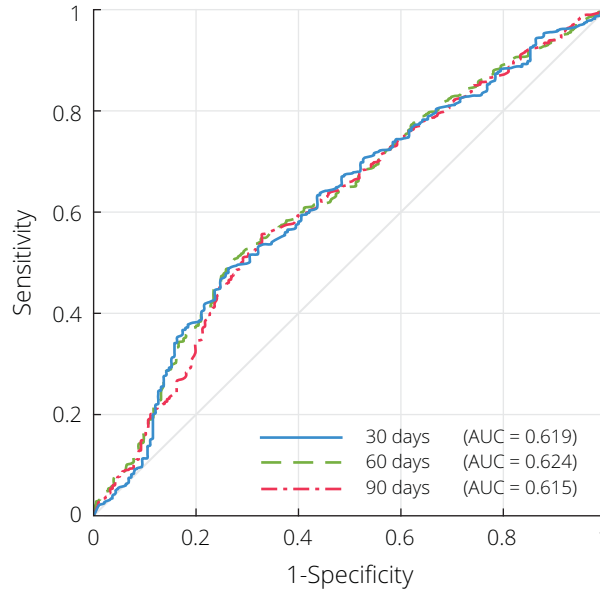


Figure 4. Time windows. ROC curves for variations of three different time windows (30, 60 and 90 days) where all the first days in each month had been excluded.

3.3 P-value approximation

When considering the performance of the p-value approximation, the dataset with emerging signals was used in the variation where all first days in each month had been excluded and where the events occurred within the first 60 days. This resulted in 2 416 combinations, where 162 were positive controls and 2 254 were negative controls. The approximation of the p-value gave a performance close to random chance (Fig. 5) with an AUC of 0.518 and is therefore not a good predictor for this type of data.

When the significance level was set to 0.05, 24 true positives were obtained for this variation of the dataset, i.e. their distributions were significantly different. The empirical distributions of TTO values for two of these true positives were plotted for visual evaluation (Fig.6). The combination Aripiprazole and Hypertension had an approximated p-value of 0.0005, while the vigiRank value was 0.087 for the same

combination. Eptifibatide together with Pulmonary haemorrhage had an approximated p-value of 0.0071 while the vigiRank value was 0.074. Based on the vigiRank values alone, these combinations would likely not be highlighted as potential signals for further in-depth clinical assessment. Furthermore, the first combination had a negative IC value of -1.478, meaning that it probably would not have been highlighted by disproportionality analysis either. However, the p-values are much lower than 0.05 for both combinations, suggesting that the associations might be of interest.

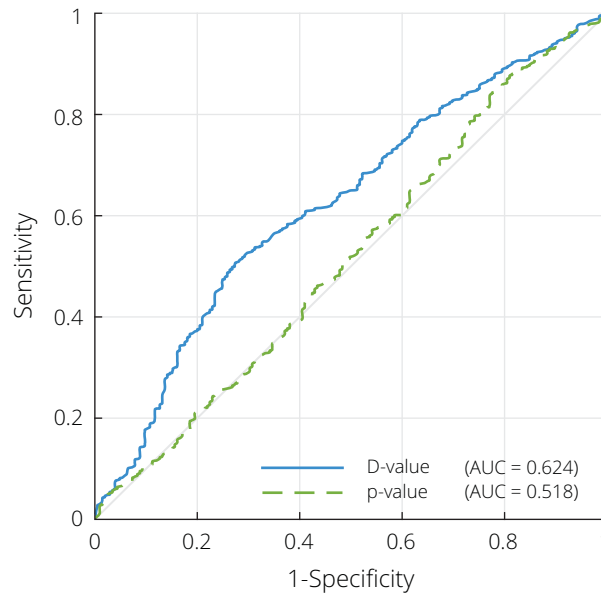


Figure 5. p-value approximation. ROC curves for the KS test result based on the D-value and the p-value of the dataset containing emerging signals. All first days in each month had been excluded and only reports containing events which occurred within the first 60 days were considered.

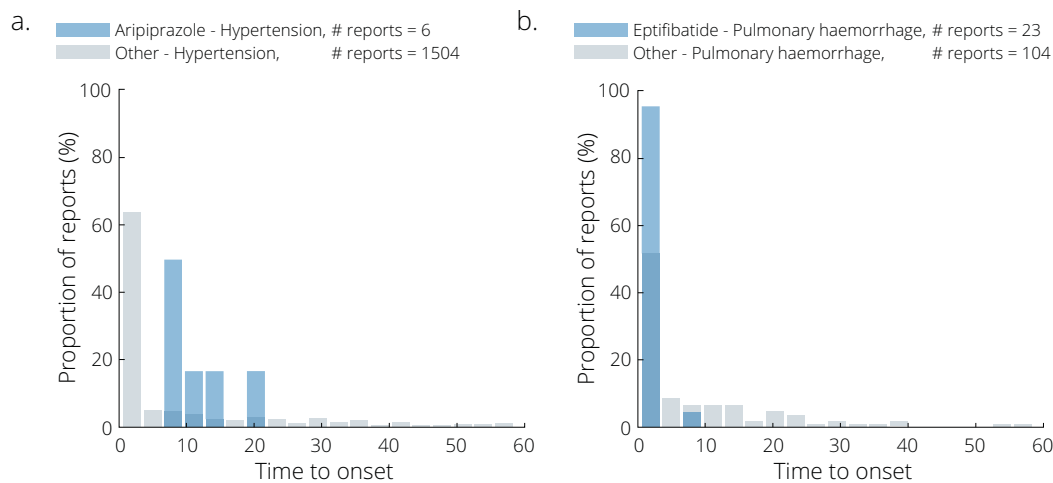


Figure 6. Distributions of true positives. Distribution plots describing two true positive combinations when the significance level was set to 0.05 for the dataset with emerging signals. **a.** Distribution of TTO values for Aripiprazole and Hypertension. **b.** Distribution of TTO values for Eptifibatide and Pulmonary haemorrhage.

3.4 vigiRank

When the KS test result was included among the variables for the refitting of vigiRank, the D-values from the dataset containing emerging signals were used. The D-values were based on reports where the first days of each month were excluded and where the events occurred within the first 60 days. However, this filtration was not performed for the other 12 variables. Only combinations for which a D-value existed were included in the model fitting process, excluding almost half of the combinations used for creating the original model. The performance was plotted in a ROC curve (Fig. 7). The coefficient for the D-value variable was estimated to zero during the lasso logistic regression, meaning that the variable was not included in the final model. Instead, the model was based on the same five variables as when vigiRank was first implemented. When including all 5 544 combinations in the refitting process by setting the missing D-values to zero, the result was similar.

The refitted vigiRank model had the highest predictive power with an AUC of 0.787 followed by the IC_{025} (AUC = 0.741), the number of reports within the specified time window (AUC = 0.708) and the KS test when used alone (AUC = 0.624). No additional filtration had been performed for the IC_{025} and the number of reports variables. Furthermore, the D-value did not correlate strongly with any of the 13 variables used in the original implementation of vigiRank.

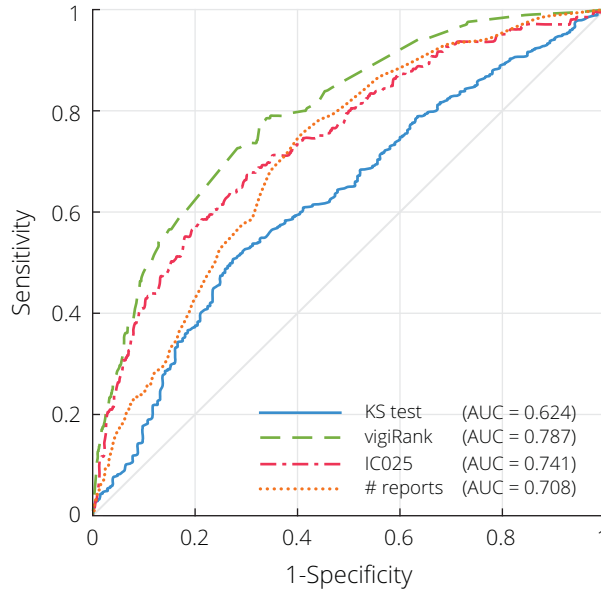


Figure 7. Comparison of predictors. ROC curves for the refitted vigiRank, IC_{025} , the number of reports and the D-value used alone as a predictor. The D-value did not contribute with enough predictive power to be included in the vigiRank model.

3.5 Prediction of established signals

Established signals were tested for comparative purposes (Fig. 8). The dataset contained signals with TTO values within the 60 days range and where all first days in each month were excluded. Of the original 31 414 combinations, 16 570 could be used for evaluation of the performance. Of these combinations, 11 551 were positive

controls and 5 019 were negative controls. Both the D-value ($AUC = 0.644$) and the p-value ($AUC = 0.622$) turned out to be better predictors for established signals than for emerging signals. Furthermore, both predictors performed better than disproportionality analysis ($AUC = 0.608$), but much worse than the raw number of reports for each combination ($AUC = 0.711$). By setting the significance level to 0.01 for the p-value predictor, a sensitivity of 0.945 and a specificity of 0.206 was obtained.

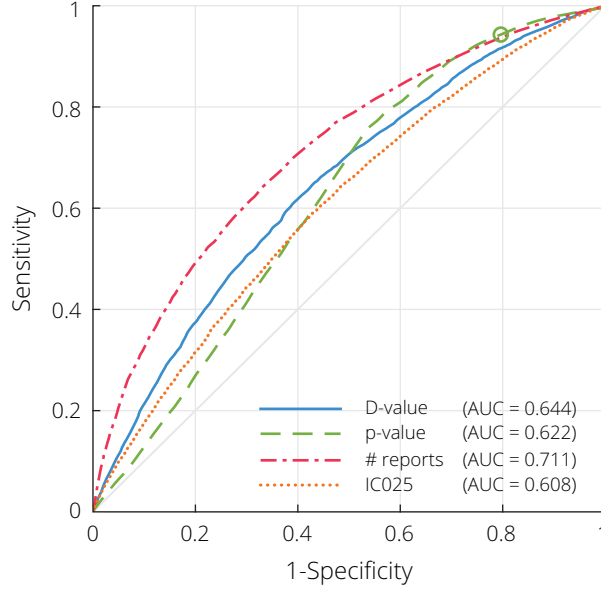


Figure 8. Established signals. ROC curves for the D-values and the p-values of the KS test, as well as for disproportionality analysis and the raw number of reports based on established signals. The dataset consisted of TTO values within the first 60 days where all first days in each month were excluded. The green circle indicates the threshold value 0.01.

4 DISCUSSION AND CONCLUSIONS

Identifying adverse drug reactions has become more important as the number of drugs and the extent of their use increases. Multiple methods have been implemented for this purpose, some more widely used than others. One method based on deviating TTO patterns has shown promising results when used on vaccine data, however, this report shows that these results might not be directly transferable to non-vaccines.

This report presents the results from an implementation of the two-sample Kolmogorov-Smirnov test in VigiBase. When the test was performed on emerging signals, the KS test statistic D did have some predictive power, while the p -value had a predictive power close to random chance. The different variations of the dataset did not affect the predictions significantly, suggesting that the overall pattern of the underlying data is a more important factor for the performance. Even though a weak predictive power could be observed using the D -value, the performance was significantly lower than that of *vigiRank* and disproportionality analysis. When performing the KS test on established signals, the performance of the predictor based on the D -value was only slightly better than the one obtained for emerging signals. However, the p -value had a predictive power significantly higher than random chance. This might be due to the larger number of reports used in the KS test for established signals compared to the much smaller number of reports used when evaluating emerging signals. The p -value approximation does not perform well when sample sizes are small, which was the case for the dataset containing emerging signals. In addition, the larger number of reports available for established signals makes the data more balanced, and hence, makes the p -value approximation more accurate.

By adding the KS test result for emerging signals to the set of variables used for the refitting of *vigiRank*, possible synergistic effects could be evaluated. However, the KS test result did not contribute to enough predictive power for it to be included in the final model, and did not seem to affect the selection of other variables. Nevertheless, the D -value was not strongly correlated with any other variable, implying that it does measure a feature not covered in the current implementation of *vigiRank*. Because of the weak predictive power, the KS test might not be the most appropriate approach of extracting information from TTO data of non-vaccines in VigiBase.

One important limitation of the KS test when used on ICSR data is the preselection of reports having a higher level of completeness. This is a result of excluding reports lacking TTO data from the calculation of the KS test statistic. The exclusion will indirectly affect which combinations that will be evaluated, which in this case will be combinations with a higher general level of completeness. Since incomplete reports

are common in ICSR databases, and since the completeness does have predictive power on its own, this must be taken into account when evaluating the method. However, this is an intrinsic limitation of the method since TTO information is needed to perform the test, making it useless for reports lacking TTO data.

An additional limitation of the results presented here is the number of reports needed for constructing an accurate distribution. The results are based on combinations represented by at least two reports. However, two reports might be too few to obtain reliable results. To avoid this issue, the lower limit for the required number of reports may be raised. However, setting a lower limit is not possible in reality and would therefore give a misleading result. Furthermore, the number of combinations fulfilling the criteria would be decreasing rapidly as the lower limit increases which leads to a greater uncertainty in the final results. This also means that raising the lower limit will not be a solution for the heavily imbalanced data in VigiBase. Most combinations in VigiBase are only represented by a few number of reports with TTO data. However, the number of reports containing only a specific drug is usually significantly higher. This leads to imbalanced data which is known to decrease the performance of the test, showing an additional limitation caused by the underlying data alone.

Earlier studies have shown positive results when applying the KS test to vaccine data, suggesting that the KS test may work as a complement to disproportionality analysis [5]. However, no studies have yet been published on the performance of the KS test on non-vaccine data. This report presents an implementation of the KS test on non-vaccines in VigiBase and also investigates performance based on the KS test statistic, the D-value, something that has not been considered earlier.

The results for established signals presented here are in line with the results presented by Van Holle et al. [5]. Based on established signals, the KS test performs better than disproportionality analysis for both vaccines and non-vaccines. However, when the raw number of reports is used as a comparative predictor, it outperforms the KS test. The same effects has been shown earlier by Norén et al. [23] and indicate that the results cannot be trusted. Instead, emerging signals should be used for more reliable results. The predictive power of the KS test is significantly lower than that of disproportionality analysis and vigiRank when evaluated on emerging signals, indicating that the KS test is not an appropriate predictor of ADRs for non-vaccine drugs. However, the impact of established signals vs. emerging signals on this type of statistical test should be further explored, e.g. by performing the KS test on vaccines using emerging signals to investigate the consistency of the results presented here.

Another suggestion for future studies is to investigate the differences between vaccines and non-vaccines. It is not obvious that the results for vaccines are directly transferable to non-vaccines because of intrinsic effects (e.g. immunological effects) and different reporting patterns. Furthermore, most vaccines are only given once or twice, making sure that the event onset is not the cause of drug accumulation during a longer period of time. To the contrary, non-vaccines are usually given during some period, which could result in a delayed TTO, possibly related to dosage. The dosage might not be the same during the whole treatment period and even if it was, the concentration of the drug at the site of action might vary depending on disease

progress. These factors among others, could result in less distinct TTO patterns for non-vaccines.

In conclusion, the two-sample Kolmogorov-Smirnov test did not perform to a satisfactory level when used alone on non-vaccine drugs. Furthermore, it did not outperform any of the currently implemented methods, i.e. disproportionality analysis and *vigiRank*. The test result did not give sufficiently independent predictive power to be included as a variable in the predictive model *vigiRank*, hindering any possible synergistic effects to take place. The two-sample Kolmogorov-Smirnov test does therefore not seem to be an appropriate method for extracting information from the TTO data for non-vaccine drugs in *VigiBase*.

5 ACKNOWLEDGEMENTS

First, I would like to give a special thanks to my supervisor Ola Caster for the opportunity to do this Master's thesis project at the UMC and for his support throughout the course of the project. I would also like to thank Kristina Juhlin for all of the help and knowledge shared regarding the databases at the UMC. Thirdly, I would like to thank Lionel Van Holle for sharing his thoughts about various aspects of the implementation and results. Finally, a big thank to everyone at the UMC, especially to the research section, for sharing ideas and suggestions about how to move forward and what to think about when dealing with various issues that have arisen during the course of the project.

6 REFERENCES

- [1] Lazarou J, Pomeranz BH, and Corey PN, "Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies," *JAMA*, vol. 279, no. 15, pp. 1200–1205, Apr. 1998.
- [2] F. R. Ernst and A. J. Grizzle, "Drug-related morbidity and mortality: updating the cost-of-illness model," *J. Am. Pharm. Assoc. WashingtonDC* 1996, vol. 41, no. 2, pp. 192–199, Apr. 2001.
- [3] D. W. Kaufman, J. P. Kelly, L. Rosenberg, T. E. Anderson, and A. A. Mitchell, "Recent patterns of medication use in the ambulatory adult population of the United States: the Slone survey," *JAMA*, vol. 287, no. 3, pp. 337–344, Jan. 2002.
- [4] S. M. Schappert and C. W. Burt, "Ambulatory care visits to physician offices, hospital outpatient departments, and emergency departments: United States, 2001-02," *Vital Health Stat. 13.*, no. 159, pp. 1–66, Feb. 2006.
- [5] L. Van Holle, Z. Zeinoun, V. Bauchau, and T. Verstraeten, "Using time-to-onset for detecting safety signals in spontaneous reports of adverse events following immunization: a proof of concept study," *Pharmacoepidemiol. Drug Saf.*, vol. 21, no. 6, pp. 603–610, 2012.
- [6] O. Caster, K. Juhlin, S. Watson, and G. N. Norén, "Improved statistical signal detection in pharmacovigilance by combining multiple strength-of-evidence aspects in vigiRank," *Drug Saf.*, vol. 37, no. 8, pp. 617–628, Aug. 2014.
- [7] World Health Organization, *The Importance of Pharmacovigilance - Safety Monitoring of Medicinal Products*. World Health Organization, 2002.
- [8] J. Talbot and J. K. Aronson, Eds., *Stephens' Detection and Evaluation of Adverse Drug Reactions: Principles and Practice*, 6 edition. Chichester, West Sussex, UK: Wiley-Blackwell, 2011.
- [9] J. Lexchin, "Why are there deadly drugs?," *BMC Med.*, vol. 13, no. 1, p. 27, Feb. 2015.
- [10] P. Waller, *An Introduction to Pharmacovigilance*, 1st ed. Hoboken: Wiley, 2009.
- [11] L. Hazell and S. A. W. Shakir, "Under-reporting of adverse drug reactions : a systematic review," *Drug Saf.*, vol. 29, no. 5, pp. 385–396, 2006.

- [12] J. Venulet and M. Helling-Borda, “WHO’s International Drug Monitoring - The Formative Years, 1968–1975: Preparatory, Pilot and Early Operational Phases,” *Drug Saf.*, vol. 33, no. 7, pp. e1–e23, Jul. 2010.
- [13] M. Lindquist, “VigiBase, the WHO Global ICSR Database System: Basic Facts,” *Drug Inf. J.*, vol. 42, no. 5, pp. 409–419, Sep. 2008.
- [14] G. N. Norén, R. Orre, A. Bate, and I. R. Edwards, “Duplicate detection in adverse drug reaction surveillance,” *Data Min. Knowl. Discov.*, vol. 14, no. 3, pp. 305–328, Feb. 2007.
- [15] A. Bate, M. Lindquist, I. R. Edwards, S. Olsson, R. Orre, A. Lansner, and R. M. De Freitas, “A Bayesian neural network method for adverse drug reaction signal generation,” *Eur. J. Clin. Pharmacol.*, vol. 54, no. 4, pp. 315–321, Jun. 1998.
- [16] G. N. Norén, J. Hopstadius, and A. Bate, “Shrinkage observed-to-expected ratios for robust and transparent large-scale pattern discovery,” *Stat. Methods Med. Res.*, vol. 22, no. 1, pp. 57–69, Feb. 2013.
- [17] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” *J. R. Stat. Soc. Ser. B Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996.
- [18] T. Bergvall, G. N. Norén, and M. Lindquist, “vigiGrade: A Tool to Identify Well-Documented Individual Case Reports and Highlight Systematic Data Quality Issues,” *Drug Saf.*, vol. 37, no. 1, pp. 65–77, Dec. 2013.
- [19] J. Hopstadius and G. N. Norén, “Robust Discovery of Local Patterns: Subsets and Stratification in Adverse Drug Reaction Surveillance,” in *Proceedings of the 2Nd ACM SIGHIT International Health Informatics Symposium*, New York, NY, USA, 2012, pp. 265–274.
- [20] L. Van Holle and V. Bauchau, “Use of Logistic Regression to Combine Two Causality Criteria for Signal Detection in Vaccine Spontaneous Report Data,” *Drug Saf.*, vol. 37, no. 12, pp. 1047–1057, Nov. 2014.
- [21] L. van Holle and V. Bauchau, “Signal detection on spontaneous reports of adverse events following immunisation: a comparison of the performance of a disproportionality-based algorithm and a time-to-onset-based algorithm,” *Pharmacoepidemiol. Drug Saf.*, vol. 23, no. 2, pp. 178–185, Feb. 2014.
- [22] Y. Alvarez, A. Hidalgo, F. Maignen, and J. Slattery, “Validation of statistical signal detection procedures in eudravigilance post-authorization data: a retrospective evaluation of the potential for earlier signalling,” *Drug Saf.*, vol. 33, no. 6, pp. 475–487, Jun. 2010.
- [23] G. N. Norén, O. Caster, K. Juhlin, and M. Lindquist, “Zoo or savannah? Choice of training ground for evidence-based pharmacovigilance,” *Drug Saf.*, vol. 37, no. 9, pp. 655–659, Sep. 2014.
- [24] A. N. Kolmogorov, “Sulla determinazione empirica di una legge di distribuzione,” *G. Dell’Istituto Ital. Degli Attuari*, vol. 4, pp. 83–91, 1933.
- [25] N. V. Smirnov, “Estimate of deviation between empirical distribution functions in two independent samples,” *Mosc. Univ. Math. Bull.*, vol. 2, no. 2, pp. 3–16, 1939.

- [26] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, 3rd ed. Chapman and Hall/CRC, 2003.
- [27] G. W. Corder and D. I. Foreman, *Nonparametric Statistics: A Step-By-Step Approach*, 2nd ed. John Wiley & Sons, 2014.
- [28] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. Cambridge University Press, 2007.
- [29] G. L. Tietjen, D. K. Kahaner, R. J. Beckman, W. J. Kennedy, and H. A. David, *Selected tables in mathematical statistics*, vol. 5. American Mathematical Soc., 1977.
- [30] R. Simard and P. L'Ecuyer, "Computing the Two-Sided Kolmogorov-Smirnov Distribution," *J. Stat. Softw.*, vol. 39, no. 11, Mar. 2011.
- [31] R Core Team, *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- [32] T. A. Arnold and J. W. Emerson, "Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions," *R J.*, vol. 3, no. 2, pp. 34–39, 2011.
- [33] SAS Institute Inc, "The NPAR1WAY procedure," in *SAS/STAT® 13.2 User's Guide*, Cary, North Carolina, 2014.
- [34] D. J. Steinskog, D. B. Tjøstheim, and N. G. Kvamstø, "A Cautionary Note on the Use of the Kolmogorov–Smirnov Test for Normality," *Mon. Weather Rev.*, vol. 135, no. 3, pp. 1151–1157, Mar. 2007.
- [35] L. B. K. Alexander Y. Gordon, "On a paradoxical property of the Kolmogorov–Smirnov two-sample test," in *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková*, Beachwood, Ohio, USA, 2010, pp. 70–74.
- [36] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

7 APPENDICES

Appendix A

```
-- =====
-- Name:          KS_DataExtraction
-- Author:        Caroline Wörn
-- Create date:   2015-03-09
--
-- Description:   Extracts data from Vigisearch2012 and prepares
--               it for the stored procedure KS_test.
--
--               For a set of drug-ADR pairs, reports containing
--               either the drug or ADR is extracted. Dates for
--               administration and ADR onset is then retrieved,
--               while removing drugs listed as concomitants as well
--               as incomplete and invalid dates. Suspected
--               duplicates are also removed.
--
--               The first day in each month can be handled in
--               three different ways. They can be kept as entered
--               into the database system, they can be removed and
--               they can be changed to the 15th of the same month.
--               When this decision has been made, time-to-onset is
--               computed as the number of days between and
--               including the start and onset date.
--
--               When multiple prescriptions for the same
--               combination is present on the same report, only
--               the one with the shortest calculated TTO value is
--               kept. The resultset is then returned.
--
-- Input:         - Table containing two columns, one for Drug ID and
--               and one for ADR ID.
--               Type: KS_COMBINATION
--               - Parameter describing the size of the time
--               window in which reports are retained.
--               Type: INT
--               - A keyword describing how the first day in a month
--               is treated. 'keep01' keeps all first days,
--               'remove01' removes all first days and 'change01'
--               changes all first days to the 15th of the same
--               month.
--               Type: NVARCHAR
--
-- Output:        - Table containing columns for Drug ID, ADR ID and
--               calculated TTO.
--
-- =====

CREATE PROCEDURE carolinew.KS_DataExtraction
(
    @Combinations AS CAROLINEW.KS_COMBINATION READONLY,
    @TimeWindow   AS INT,
    @FirstDay     AS NVARCHAR(10)
)
AS
BEGIN
```

```

SET NOCOUNT ON;
SET ANSI_WARNINGS OFF;

-- Select all combinations including either the drug or the ADR from
-- the set of combinations
SELECT DISTINCT
    AD.VmDrug_Id,
    AD.VmAdrTerm_Id,
    AD.Report_Id
INTO
    #CombinationsData
FROM
    VM_NewTriages_2.VmDrugAdr AD
WHERE
    Ad.VmAdrTerm_Id IN (SELECT AdrTerm_Id FROM @Combinations) OR
    Ad.VmDrug_Id IN (SELECT Drug_Id FROM @Combinations);

-- Retrieve DateStart and DateOnset from reports listed as suspected
-- or interacting (Basis = 1,3) and which are not suspected
-- duplicates (Stratum_Id = 2)
SELECT DISTINCT
    T.VmDrug_Id,
    T.VmAdrTerm_Id,
    T.Report_Id,
    D.DateStart,
    A.DateOnset
INTO
    #RawData
FROM
    #CombinationsData T
    INNER JOIN Vigisearch2012.VS_2.Patient P
        ON P.Report_Id = T.Report_Id
    INNER JOIN Vigiminedda.VM_NewTriages_2.VmAdrTermAdrMapping_link AL
        ON AL.VmAdrTerm_Id = T.VmAdrTerm_Id
    INNER JOIN Vigisearch2012.VS_2.MappedReportedTerm MT
        ON AL.AdrMapping_Id = MT.AdrMapping_ID
    INNER JOIN Vigisearch2012.VS_2.ADR A
        ON MT.MappedReportedTermID = A.MappedReportedTermID AND
        A.Patient_Id = P.Patient_Id

    INNER JOIN Vigiminedda.VM_NewTriages_2.VmDrugMedProd_link DL
        ON DL.VmDrug_Id = T.VmDrug_Id
    INNER JOIN Vigisearch2012.VS_2.MappedReportedDrug MD
        ON DL.MedicinalProd_Id = MD.MedicinalProductID
    INNER JOIN Vigisearch2012.VS_2.Drug D
        ON D.MappedReportedDrugID = Md.MappedReportedDrugID AND
        D.Patient_Id = P.Patient_Id AND D.Basis IN ('1','3')

    INNER JOIN Vigiminedda.VM_NewTriages_2.VmReportsStrata RS
        ON T.Report_Id = RS.Report_Id
WHERE
    (RS.Stratum_Id = 2);

-- Select only complete and valid dates
SELECT *
INTO
    #ValidDates
FROM

```

```

#RawData
WHERE
    (LEN(DateStart) = 8 AND LEN(DateOnset) = 8) AND
    (ISDATE(DateStart) = 1 AND ISDATE(DateOnset) = 1);

-- Handle the first day in a month
CREATE TABLE #TTORawData
(
    VmDrug_Id      INT,
    VmAdrTerm_Id   SMALLINT,
    Report_Id      INT,
    DateStart      DATE,
    DateOnset      DATE
)
IF @FirstDay = 'keep01'
BEGIN
    INSERT INTO
        #TTORawData
    SELECT
        VmDrug_Id,
        VmAdrTerm_Id,
        Report_Id,
        CAST(DateStart AS DATE) AS DateStart,
        CAST(DateOnset AS DATE) AS DateOnset
    FROM #ValidDates
END
ELSE IF @FirstDay = 'remove01'
BEGIN
    INSERT INTO
        #TTORawData
    SELECT
        VmDrug_Id,
        VmAdrTerm_Id,
        Report_Id,
        CAST(DateStart AS DATE) AS DateStart,
        CAST(DateOnset AS DATE) AS DateOnset
    FROM #ValidDates
    WHERE
        RIGHT(DateStart,2) <> '01'
        AND RIGHT(DateOnset,2) <> '01'
END
ELSE IF @FirstDay = 'change01'
BEGIN
    INSERT INTO
        #TTORawData
    SELECT
        VmDrug_Id,
        VmAdrTerm_Id,
        Report_Id,
        CAST(DateStart AS DATE) AS DateStart,
        CAST(DateOnset AS DATE) AS DateOnset
    FROM #ValidDates
    WHERE
        RIGHT(DateStart,2) <> '01' AND
        RIGHT(DateOnset,2) <> '01'
    UNION ALL
    SELECT
        VmDrug_Id,
        VmAdrTerm_Id,
        Report_Id,

```

```

        CAST(DATEADD(DAY, 14, DateStart) AS DATE) AS DateStart,
        CAST(DateOnset AS DATE) AS DateOnset
FROM #ValidDates
WHERE
    RIGHT(DateStart,2) = '01' AND
    RIGHT(DateOnset,2) <> '01'
UNION ALL
SELECT
    VmDrug_Id,
    VmAdrTerm_Id,
    Report_Id,
    CAST(DateStart AS DATE),
    CAST(DATEADD(DAY, 14, DateOnset) AS DATE) AS DateOnset
FROM #ValidDates
WHERE
    RIGHT(DateStart,2) <> '01' AND
    RIGHT(DateOnset,2) = '01'
UNION ALL
SELECT
    VmDrug_Id,
    VmAdrTerm_Id,
    Report_Id,
    CAST(DATEADD(DAY, 14, DateStart) AS DATE) AS DateStart,
    CAST(DATEADD(DAY, 14, DateOnset) AS DATE) AS DateOnset
FROM #ValidDates
WHERE
    RIGHT(DateStart,2) = '01' AND
    RIGHT(DateOnset,2) = '01'

END;

-- Compute TTO and select reports within specified time window
SELECT
    VmDrug_Id,
    VmAdrTerm_Id,
    Report_Id,
    DateStart,
    DateOnset,
    DATEDIFF(DAY, DateStart, DateOnset) AS TTO
INTO
    #TTOFinalData
FROM
    #TTORawData
WHERE
    DATEDIFF(DAY, DateStart, DateOnset) BETWEEN 0 AND @TimeWindow;

-- When multiple instances are present, select the one with
--   shortest TTO
SELECT T1.*
INTO
    #TTODataMod
FROM
    (
        SELECT
            VmDrug_Id,
            VmAdrTerm_Id,
            Report_Id,
            MIN(TTO) AS MinTTO
        FROM
            #TTOFinalData

```

```
        GROUP BY
            Report_Id,
            VmDrug_Id,
            VmAdrTerm_Id
    ) AS T2
INNER JOIN #TTOFinalData T1
    ON
        T2.Report_Id = T1.Report_Id AND
        T2.MinTTO = T1.TTO AND
        T2.VmDrug_Id = T1.VmDrug_Id AND
        T2.VmAdrTerm_ID = T1.VmAdrTerm_Id;

-- Return result set
SELECT
    VmDrug_Id,
    VmAdrTerm_Id,
    TTO
FROM
    #TTODataMod;

-- Clean up temporary tables
DROP TABLE #CombinationsData;
DROP TABLE #RawData;
DROP TABLE #ValidDates;
DROP TABLE #TTORawData;
DROP TABLE #TTOFinalData;
DROP TABLE #TTODataMod;

END
```

Appendix B

```
-- =====
-- Name:          KS_Test
-- Author:        Caroline Wörn
-- Create date:   2015-02-20
-- Description:   Performs a two-sample Kolmogorov-Smirnov test.
--               Computes the KS test statistic D, p-value and
--               hypothesis based on significance level. The null
--               hypothesis (H0) is that the two sample sets originate
--               from the same distribution.
--
-- Input:         - Table containing instances of specified Drug ID and ADR
--               ID as well as their computed time-to-onsets
--               Type: KS_TTODATA
--               - Table containing combinations to be tested.
--               Type: KS_COMBINATION
--               - Significance level. If no significance level is
--               specified, the default level 0.05 is used.
--               Type: DECIMAL
--               - Type of test. Specified by the string 'Between
--               Events' or 'Between Drugs'.
--               Type: NVARCHAR
--
-- Output:        - Result set containing combinations, D-values,
--               p-values, hypotheses and information about the
--               number of reports in each sample group.
--               (1 = reject H0, 0 = cannot reject H0)
--
-- =====
CREATE PROCEDURE carolinew.KS_Test
(
    @TTOData          AS CAROLINEW.KS_TTODATA          READONLY,
    @Combinations     AS CAROLINEW.KS_COMBINATION     READONLY,
    @Alpha            AS DECIMAL(18,4) = 0.05,
    @TestType         AS NVARCHAR(15)
)
AS
BEGIN

    SET NOCOUNT ON;
    SET ANSI_WARNINGS OFF;

    DECLARE
        @RowCounter    INT = 1;

    -- Generate values 1-101 for approximation of null hypothesis
    -- distribution
    SELECT TOP (101) j = ROW_NUMBER() OVER (ORDER BY [object_id])
    INTO #J
    FROM sys.all_objects
    ORDER BY j;

    -- Create table to hold resulting test statistics
    CREATE TABLE #TestStatistics
```

```

(
    RowCounter          INT IDENTITY(1,1),
    Drug_Id             INT,
    AdrTerm_Id          SMALLINT,
    KSstatistic         DECIMAL(18,4),
    Pvalue              DECIMAL(18,4),
    Hypothesis          BIT,
    TooFewReports       VARCHAR(30),
    NoCompDist          VARCHAR(30),
    NumInDist           INT,
    NumInCompDist       INT
);

-- Insert all unique combinations to test
INSERT INTO #TestStatistics
(
    Drug_Id,
    AdrTerm_Id
)
SELECT DISTINCT
    Drug_Id,
    AdrTerm_Id
FROM
    @Combinations;

-- Loop through all unique combinations
WHILE ((SELECT MAX(RowCounter) FROM #TestStatistics) >= @RowCounter)
BEGIN

    DECLARE
        @Drug_Id          INT,
        @AdrTerm_Id       SMALLINT,
        @N1               DECIMAL(18,4),
        @N2               DECIMAL(18,4),
        @N                DECIMAL(18,4),
        @KSstatistic      DECIMAL(18,4),
        @Lambda           DECIMAL(18,4),
        @Pvalue           DECIMAL(18,4),
        @Hypothesis       BIT,
        @ErrorTooFew      VARCHAR(30) = NULL,
        @ErrorNoCompDist  VARCHAR(30) = NULL,
        @NumInDist        INT,
        @NumInCompDist    INT;

    -- Store current drug and ADR term ID
    SELECT
        @Drug_Id = T.Drug_Id,
        @AdrTerm_Id = T.AdrTerm_Id
    FROM
        #TestStatistics T
    WHERE
        T.RowCounter = @RowCounter;

    -- Compute empirical frequencies and cumulative sums
    WITH Frequency
    (
        TTO,

```

```

        [Frequency 1],
        [Frequency 2]
    )
AS
(
    SELECT
        TTO,
        SUM(CASE WHEN Drug_Id = @Drug_Id AND AdrTerm_Id = @AdrTerm_Id
            THEN 1 END) AS [Frequency 1],
        CASE
            WHEN @TestType = 'Between Events'
            THEN
                (SELECT SUM(CASE WHEN
                    Drug_Id = @Drug_Id AND
                    AdrTerm_Id <> @AdrTerm_Id
                    THEN 1
                    END))
            WHEN @TestType = 'Between Drugs'
            THEN
                (SELECT SUM(CASE WHEN
                    Drug_Id <> @Drug_Id AND
                    AdrTerm_Id = @AdrTerm_Id
                    THEN 1
                    END))
            END AS [Frequency 2]
        FROM @TTOData
        GROUP BY TTO
    )
SELECT
    TTO,
    [Frequency 1],
    SUM([Frequency 1]) OVER (ORDER BY TTO ROWS UNBOUNDED PRECEDING)
        AS [CumulativeSum 1],
    [Frequency 2],
    SUM([Frequency 2]) OVER (ORDER BY TTO ROWS UNBOUNDED PRECEDING)
        AS [CumulativeSum 2]
INTO #CumulativeSums
FROM Frequency
ORDER BY TTO;

-- Store the number of samples in each distribution
SELECT
    @N1 = CAST(MAX([CumulativeSum 1]) AS DECIMAL(18,4)),
    @N2 = CAST(MAX([CumulativeSum 2]) AS DECIMAL(18,4))
FROM
    #CumulativeSums;
SET @N = (SELECT (@N1 * @N2) / (@N1 + @N2));

-- Divide the cumulative sums by the number of samples in each
-- distribution
SELECT
    [CumulativeSum 1] / @N1 AS [CDF 1],
    [CumulativeSum 2] / @N2 AS [CDF 2]
INTO #CDF
FROM #CumulativeSums;

-- Convert NULL to 0 for upcoming subtraction
UPDATE #CDF

```

```

SET
    [CDF 1] = ISNULL([CDF 1], 0),
    [CDF 2] = ISNULL([CDF 2], 0);

-- Compute 2-sided test statistic: D = max|F1(x) - F2(x)|
SET @KSstatistic =
    (SELECT MAX (ABS ([CDF 1] - [CDF 2]))
     FROM #CDF);

-- MATLAB's p-value approximation
/*
-- Compute approximation of the distribution for null hypothesis
SET @Lambda =
    (SELECT (SQRT (@N) + 0.12 + 0.11 / SQRT (@N)) * @KSstatistic);
-- Compute the asymptotic distribution (Q-function) to approximate
-- the 2-sided p-value
SET @Pvalue =
    (SELECT 2 * SUM (POWER ((-1), j-1) * EXP ((-2) *
        POWER(@Lambda, 2) * POWER(j, 2))) FROM #J);
*/

-- SAS's p-value approximation
SET @Lambda =
    (SELECT (SQRT (@N)) * @KSstatistic);
-- Compute the asymptotic distribution (Q-function) to approximate
-- the 2-sided p-value
SET @Pvalue =
    (SELECT 2 * SUM (POWER ((-1), j-1) * EXP ((-2) *
        POWER(@Lambda, 2) * POWER(j, 2))) FROM #J);

-- Nonparametric Statistics: A Step-by-Step Approach, 2nd Edition
-- (p-value approximation)
/*
SET @Lambda =
    (SELECT (SQRT (@N)) * @KSstatistic);

-- Compute the asymptotic distribution (Q-function) to approximate
-- the 2-sided p-value
DECLARE @Q DECIMAL(18,4);

IF (0 <= @Lambda) AND (@Lambda < 0.27)
    BEGIN
        SET @Pvalue = 1;
    END
ELSE IF (0.27 <= @Lambda) AND (@Lambda < 1)
    BEGIN
        SET @Q = EXP ((-1.233701) * POWER (@Lambda, (-2)))
        SET @Pvalue = 1 - (2.506628 / @Lambda) * (@Q + POWER (@Q, 9)
            + POWER (@Q, 25));
    END
ELSE IF (1 <= @Lambda) AND (@Lambda < 3.1)
    BEGIN
        SET @Q = EXP ((-2) * POWER (@Lambda, 2))
        SET @Pvalue = 2 * (@Q - POWER (@Q, 4) + POWER (@Q, 9)
            - POWER (@Q, 16));
    END
ELSE IF (3.1 <= @Lambda)

```

```

        BEGIN
            SET @Pvalue = 0;
        END
    */

-- Determine if hypothesis holds
IF @Alpha >= @Pvalue
    SET @Hypothesis = 1
ELSE
    SET @Hypothesis = 0;

-- Error Information:
-- Find combinations that do not occur frequently enough to reject
-- null hypothesis
IF NOT EXISTS
(
    SELECT Drug_Id, AdrTerm_Id
    FROM @TTOData
    WHERE AdrTerm_Id = @AdrTerm_Id AND Drug_Id = @Drug_Id
    GROUP BY Drug_Id, AdrTerm_Id
    HAVING COUNT(*) > 1
)
BEGIN
    SET @ErrorTooFew = 'Too few reports'
END;

-- Error Information:
-- Find combinations that have nothing to compare with
IF @TestType = 'Between Drugs' AND NOT EXISTS
(
    SELECT AdrTerm_Id
    FROM @TTOData
    WHERE AdrTerm_Id = @AdrTerm_Id AND Drug_Id <> @Drug_Id
    GROUP BY AdrTerm_Id
    HAVING COUNT(*) > 1
)
BEGIN
    SET @ErrorNoCompDist = 'No comparative distribution'
END
ELSE IF @TestType = 'Between Events' AND NOT EXISTS
(
    SELECT Drug_Id
    FROM @TTOData
    WHERE Drug_Id = @Drug_Id AND AdrTerm_Id <> @AdrTerm_Id
    GROUP BY Drug_Id
    HAVING COUNT(*) > 1
)
BEGIN
    SET @ErrorNoCompDist = 'No comparative distribution'
END;

-- Number in distribution
SET @NumInDist = (SELECT COUNT(Drug_Id)
                  FROM @TTOData
                  WHERE AdrTerm_Id = @AdrTerm_Id AND
                        Drug_Id = @Drug_Id
                  GROUP BY Drug_Id, AdrTerm_Id)

```

```

-- Number in Comparative distribution
IF @TestType = 'Between Drugs'
BEGIN
    SET @NumInCompDist = (SELECT COUNT(AdrTerm_Id)
                          FROM @TTOData
                          WHERE AdrTerm_Id = @AdrTerm_Id AND
                                Drug_Id <> @Drug_Id
                          GROUP BY AdrTerm_Id)
END
ELSE IF @TestType = 'Between Events'
BEGIN
    SET @NumInCompDist = (SELECT COUNT(Drug_Id)
                          FROM @TTOData
                          WHERE AdrTerm_Id <> @AdrTerm_Id AND
                                Drug_Id = @Drug_Id
                          GROUP BY Drug_Id)
END

-- Collect test statistics
UPDATE T
SET
    T.Hypothesis = @Hypothesis,
    T.Pvalue = @Pvalue,
    T.KSstatistic = @KSstatistic,
    T.TooFewReports = @ErrorTooFew,
    T.NoCompDist = @ErrorNoCompDist,
    T.NumInDist = @NumInDist,
    T.NumInCompDist = @NumInCompDist
FROM #TestStatistics AS T
WHERE
    T.Drug_Id = @Drug_Id AND
    T.AdrTerm_Id = @AdrTerm_Id;

-- Update loop counter
SET @RowCounter = @RowCounter + 1;

-- Drop tables
DROP TABLE #CumulativeSums;
DROP TABLE #CDF;

END -- End while-loop

-- Return test statistics
SELECT
    Drug_Id,
    AdrTerm_Id,
    KSstatistic,
    Pvalue,
    Hypothesis,
    TooFewReports,
    NoCompDist,
    NumInDist,
    NumInCompDist
FROM
    #TestStatistics;

```

```
-- Clean up temporary tables
DROP TABLE #J;
DROP TABLE #TestStatistics;

END
```

Appendix C

```
-- =====
-- Name:          KS_Performance
-- Author:        Caroline Wörn
-- Create date:   2015-02-20
-- Description:   Computes Specificity and Sensitivity for the result
--               set from the KS_test procedure.
--
-- Input:         - Table containing result set from KS_Test.
--               Type: KS_HYPOTHESIS
--               - Table containing unique positive controls.
--               Type: KS_COMBINATION
--               - Table containing unique negative controls.
--               Type: KS_COMBINATION
--
-- Output:        - Table containing performance measures.
--
-- =====
CREATE PROCEDURE carolinew.KS_Performance
(
    @CombinationHypothesis AS CAROLINEW.KS_HYPOTHESIS    READONLY,
    @PositiveCombinations AS CAROLINEW.KS_COMBINATION    READONLY,
    @NegativeCombinations AS CAROLINEW.KS_COMBINATION    READONLY
)
AS
BEGIN

    SET NOCOUNT ON;

    -- Create table to hold results
    CREATE TABLE #PerformanceData
    (
        Drug_Id          INT          NOT NULL,
        AdrTerm_Id       INT          NOT NULL,
        KSStatistic       DECIMAL(18,4) NULL,
        Pvalue           DECIMAL(18,4) NULL,
        Hypothesis        BIT          NULL,
        TrueOutcome       BIT          NULL,
        ErrorType         SMALLINT     NULL,
        TooFewReports     VARCHAR(30)  NULL,
        NoCompDist        VARCHAR(30)  NULL,
        NumInDist         INT          NULL,
        NumInCompDist     INT          NULL
    );

    -- Populate table with initial values
    INSERT INTO #PerformanceData
    (
        Drug_Id,
        AdrTerm_Id,
        KSStatistic,
        Pvalue,
        Hypothesis,
        TooFewReports,
        NoCompDist,
        NumInDist,

```

```

        NumInCompDist
    )
SELECT
    Drug_Id,
    AdrTerm_Id,
    KSStatistic,
    Pvalue,
    Hypothesis,
    TooFewReports,
    NoCompDist,
    NumInDist,
    NumInCompDist
FROM
    @CombinationHypothesis

-- Add true outcomes in table PerformanceData
UPDATE PD
SET PD.TrueOutcome = 1
FROM
    #PerformanceData AS PD
INNER JOIN
    @PositiveCombinations AS PC
ON PD.Drug_Id = PC.Drug_Id
AND PD.AdrTerm_Id = PC.AdrTerm_Id
WHERE
    PD.Drug_Id = PC.Drug_Id
    AND PD.AdrTerm_Id = PC.AdrTerm_Id;

UPDATE PD
SET PD.TrueOutcome = 0
FROM
    #PerformanceData AS PD
INNER JOIN
    @NegativeCombinations AS NC
ON PD.Drug_Id = NC.Drug_Id
AND PD.AdrTerm_Id = NC.AdrTerm_Id
WHERE
    PD.Drug_Id = NC.Drug_Id
    AND PD.AdrTerm_Id = NC.AdrTerm_Id;

-- Set type of error
-- 1 = True Positive
-- 2 = True Negative
-- 3 = False Positive
-- 4 = False Negative
UPDATE #PerformanceData
SET #PerformanceData.ErrorType =
(
    CASE
        WHEN (PD.Hypothesis = 1 AND PD.TrueOutcome = 1) THEN 1
        WHEN (PD.Hypothesis = 0 AND PD.TrueOutcome = 0) THEN 2
        WHEN (PD.Hypothesis = 1 AND PD.TrueOutcome = 0) THEN 3
        WHEN (PD.Hypothesis = 0 AND PD.TrueOutcome = 1) THEN 4
    END
)
FROM #PerformanceData AS PD;

```

```

-- Compute and return performance measures (sensitivity and
  specificity)
WITH Performance
(
  [# Data Points],
  [True Positives],
  [True Negatives],
  [False Positives],
  [False Negatives]
)
AS
(
  SELECT
    COUNT (CASE WHEN TrueOutcome IS NOT NULL THEN 1.0 END)
      AS [# Data Points],
    COUNT (CASE WHEN ErrorType = 1 THEN 1.0 END)
      AS [True Positives],
    COUNT (CASE WHEN ErrorType = 2 THEN 1.0 END)
      AS [True Negatives],
    COUNT (CASE WHEN ErrorType = 3 THEN 1.0 END)
      AS [False Positives],
    COUNT (CASE WHEN ErrorType = 4 THEN 1.0 END)
      AS [False Negatives]
  FROM
    #PerformanceData
)
SELECT
  [# Data Points],
  [True Positives],
  [True Negatives],
  [False Positives],
  [False Negatives],
  CASE
    WHEN (([True Negatives] + [False Positives]) <> 0 AND
      ([True Negatives] + [False Positives]) IS NOT NULL)
    THEN
      CAST((CAST ([True Negatives] AS DECIMAL(18,4)) /
        ([True Negatives] + [False Positives])) AS VARCHAR(20))
    ELSE 'Division by zero'
  END AS Specificity,
  CASE
    WHEN (([True Positives] + [False Negatives]) <> 0 AND
      ([True Positives] + [False Negatives]) IS NOT NULL)
    THEN
      CAST((CAST ([True Positives] AS DECIMAL(18,4)) /
        ([True Positives] + [False Negatives])) AS VARCHAR(20))
    ELSE 'Division by zero'
  END AS Sensitivity
FROM
  Performance

-- Return performance estimates
SELECT *
FROM
  #PerformanceData
WHERE
  TrueOutcome IS NOT NULL AND
  TooFewReports IS NOT NULL AND
  NoCompDist IS NOT NULL;

```

```
-- Clean up temporary tables
DROP TABLE #PerformanceData;

END
GO
```

Appendix D

The p-value of the two-sample Kolmogorov-Smirnov test can be approximated in multiple ways. Here, two additional approximations are presented.

The MATLAB function `kstest2()` uses an approximation which is based on the algorithm described in the book Numerical Recipes [28] and can be viewed below. The calculations are based on the effective number of values, N , where N_1 is the number of values in the first sample and N_2 is the number of values in the second sample.

$$N = \frac{N_1 N_2}{N_1 + N_2}$$

The asymptotic distribution under the null hypothesis (the critical d-value) is first approximated (λ) and then used in the calculation of the p-value. The documentation of the function claims that the result is reasonably accurate for samples such as $N \geq 4$. However, it also states that the conclusion based on the p-value approximation will not necessarily agree with the conclusion based on the KS test statistic D . Their implementation has been criticized for being inaccurate and for not providing an exact calculation for small samples sizes [30]. However, they use an approximation of the asymptotic distribution which is adjusted for faster convergence towards the true distribution.

$$\lambda = \sqrt{N} + 0.12 + \frac{0.11}{\sqrt{N}} * D$$

$$P(D > d) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 \lambda^2}$$

The second approach presented here is of a more complex type [27]. It approximates the p-value differently based on the value of λ , which is calculated as above. If $0 \leq \lambda < 0.27$, then the p-value is set to 1. If $0.27 \leq \lambda < 1$, then the p-value is approximated as below:

$$p = 1 - \frac{2.506628}{\lambda} (Q + Q^9 + Q^{25}), \text{ where } Q = e^{-1.233701\lambda^{-2}}$$

If $1 \leq \lambda < 3.1$, then the p-value is approximated as:

$$p = 2(Q - Q^4 + Q^9 - Q^{16}), \text{ where } Q = e^{-2\lambda^2} \quad (16)$$

And lastly, if $\lambda \geq 3.1$, then the p-value is set to 0.

The different approximations give very similar results on the dataset with emerging signals. This indicates that the slight changes made to the original Smirnov approximation [25] does not affect the end result significantly when used on this type of data. Furthermore, this means that focus can be put on other aspects of the KS test without the need to compare the results using multiple approximations.