Risk based monitoring in clinical studies – improving data quality

Sara K. S. Bengtsson

Bachelor's thesis
Department of Mathematical Statistics
Lund University

In collaboration with
TFS (Trial Form Support)
Lund, Sweden

Title:	Risk based monitoring in clinical studies – improving data quality
Author:	Sara K. S. Bengtsson
Supervised by:	Umberto Picchini (Lund University) Anders Nordlund (TFS) Anna Bertland (TFS)

Bachelor's thesis in mathematical statistics performed at the Department of Mathematical Statistics, Lund University, Lund, Sweden in collaboration with TFS (Trial Form Support), Lund, Sweden.

Abstract

Study quality is of paramount importance in clinical studies to ensure patient safety and reliable evaluation of the treatment, where the latter also entails the safety of future patients. Therefore, for example monitoring is required to minimise any risk of quality loss. Traditionally, the method for monitoring has been 100% source data verification which is costly and not sufficient. Today, authorities worldwide recommend risk based monitoring (RBM), which is a tool to monitor study site activities and it signals for unexpected deviations in processes or in data. RBM is to a large extent based on central statistical monitoring (CSM) using statistical analysis, and organisations involved in clinical studies are in the process of implementing these recommendations. However, since the methods are still in a stage of development further knowledge on the subject is needed. In this thesis, RBM was reviewed and selected methods were used to monitor a subsection of data from a clinical study conducted by TFS (Trial Form Support, Lund, Sweden). The key risk indicators for adverse events and serious adverse events were analysed using supervised and unsupervised statistical analysis which resulted in 19% of the sites being flagged for further investigation. Further adjustments of the methods are needed. The major difficulties in implementing RBM lie in the set-up and especially in that of supervised analysis. Further studies that share technical details and hands-on experience of CSM are needed to drive the development of RBM in clinical studies globally for general study quality improvement.

Contents

2. Introduction

2.1. Clinical studies

Phase I-IV clinical studies aim to evaluate a novel treatment paradigm, such as a potential pharmaceutical, in humans (appendix 1). Phase III studies are randomized controlled trials (RCT), these involve a large number of study subjects and are conducted at several sites and generally in several countries. By including subjects across a multitude of countries and sites, the study group is more diverse, resulting in a group that is more representative of the general population. Furthermore, the treatment groups are assumed to differ only in the administered treatment due to randomisation at site level. Therefore, any treatment effect should be replicated at all sites where the trial is conducted. However, as a consequence several medical professionals as well as several pieces of technical equipment are involved in the data collection in a given study. Hence, statistical variance can depend not only on normal variations but on study site or country. In turn, this can be caused by differences in equipment calibration, in local medical practices, climate, culture etcetera. Also, human error can be a source of data variation leading to increased statistical variance.

2.2. Study quality and risk of reduced quality

Study quality is required for two reasons; study subject safety and unbiased evaluation of the treatment, where the latter also entails the safety of future patients. This is concluded by the International Conference of Harmonization (ICH) and its Guidelines for Good Clinical Practice (GCP) (ICH E6 (R2) GCP). In the concept of study quality, emphasis can be put on different sections of a study such as study site quality, data quality, process quality, compliance to protocol and quality has to be ensured within all these sections (Timmermans 2016).

A way to ensure quality is to consider the risk of quality loss. In risk theory, risk is defined by a hazard and its potential impact. In clinical studies, the relevant hazards are classified as; design errors, procedural errors, recording errors, and analytical errors and in terms of data quality procedural and recording errors are considered (Timmermans 2016). The potential impacts were mentioned above; namely patient safety and validity of the treatment assessment. The risk of reduced quality is handled by monitoring any issues that interfere with the above quality aspects and by acting on these. Four different types of issues have been described based on their causes and the related intent behind each cause; error, sloppiness, tampering and fraud. In practice, this refers to for example technical problems, misunderstandings leading to procedural errors, sloppiness leading to data entry errors or incorrect data due to tampering (i.e. manipulation of data with or without intention to affect the study outcome).

To ensure study subject safety it is crucial to monitor adverse events (AE). AE are unexpected biological measurements or side effects identified in the study subjects (the definition of AE is given below, quoted from the ICH E6 (R2) GCP Guidelines). The severity and the frequency of AE, and that of serious AE (SAE) need to be monitored.

[An adverse event is:] *"Any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have a causal relationship with this treatment. An adverse event (AE) can therefore be any unfavourable and unintended sign (including an abnormal laboratory finding), symptom, or disease temporally associated with the use of a medicinal (investigational) product, whether or not related to the medicinal (investigational) product."*

2.3. Monitoring

Several tools are used to ensure quality in clinical studies; for example to set up an explicit strategy for a given study and ensure that all involved parties understand and agree to this strategy, and to monitor the ongoing processes to make sure ethical questions are given top priority and that the protocol is followed. Traditionally, study monitoring has been done by 100% source data verification (SDV). SDV entails for example site visits and thorough screening of the transcription between the source data and the clinical data base to find erroneous or extreme values that require evaluation and measure or correction. This is a very costly and slow method and thus not the best option for all studies. Furthermore, it has been reported that SDV may not be as effective as alternative approaches (Andersen 2014). An alternative approach to study monitoring is to use source data review (SDR) which focuses on the process behind the data transcription rather than the transcription itself. SDR combines local and centralised monitoring and focuses the efforts to where risk of quality loss is identified (TransCelerateBioPharma 2013). This reasoning is referred to as risk based monitoring (RBM) and is now recommended by the Food and Drug Administration (FDA, USA) and the European Medicines Agency (EMA, EU). RBM is performed centrally and aims to find potential issues by using for example statistical approaches to compare across countries and study sites, respectively. Globally, organisations working with clinical studies have already or have started to implement RBM and are still adopting to these guidelines. Therefore, there is a need to explore the concept of RBM further and to evaluate the methods involved to allow for efficient development of strategies.

3. Aim and Objectives

The aim of this thesis is to further the theoretical and practical understanding of RBM (Risk Based Monitoring).

The objectives of this thesis are to:

- Review the meaning of RBM in the GCP (Good Clinical Practice) context, including the concepts of CSM (Central Statistical Monitoring) and KRI (Key Risk Indicators).
- Perform a study simulating RBM to evaluate selected aspects of CSM. More specifically:
    - Compute KRI outcome for each site and/or country to evaluate their compliance, for one (or more) selected KRI(s) for which thresholds have been defined.
    - Explore suitable variable(s) to identify deviant sites and/or countries, for one (or more) selected KRI(s) for which thresholds have not been defined. If possible, propose sensible threshold levels.
    - Explore univariate and multivariate analysis of continuous variables to identify deviant sites and/or countries.

4. Review

4.1. Risk based monitoring

Risk based monitoring (RBM) is performed centrally and aims to find potential issues by comparison across countries and study sites, respectively (TransCelerateBioPharma 2013). Furthermore, the goal in using RBM is to focus efforts to where risk of quality loss is identified rather than focus on 100% SDV. It allows for analysis of data at an ongoing basis. The statistical part of RBM is usually referred to as central statistical monitoring (CSM). However, the involved terms are not used consistently and both RBM and CSM can pertain to different aspects of a general centralised monitoring depending on emphasis. See figure 1 for a proposed schematic representation of the RBM concept with a focus on statistical analysis.

RBM serves to monitor 1) the process (including project planning, data collection and analysis) and 2) the data. The methods used to ensure compliance to GCP are either statistically driven (CSM) or of other types; such as safety reports or audits. In this study, only CSM, that is, monitoring of data and of the process to collect data, will be considered since other processes are monitored using methods that are not relevant from a statistical point of view. CSM can be split into the categories supervised and unsupervised analysis (Oba 2016), which are described further below. In practise, CSM is performed at selected time points in between first patient in (FPI; meaning the date when the first patient is recruited) and database lock (the date after which no more data is entered into the study data base, the data is considered clean; meaning no outstanding queries, medical events are coded and approved and the investigator has applied his/her signature) to monitor the progress of the study (figure 2). The benefits from using an array of methods has been discussed previously and a combination of supervised and unsupervised analysis is recommended (Buyse 2014). On can also refer to KRI-based analysis which may include both supervised and unsupervised analysis. Taken together, RBM have the potential to result in a more efficient method compared to local monitoring, thereby increasing the quality in clinical studies while possibly decreasing costs (eClinical Forum 2012).

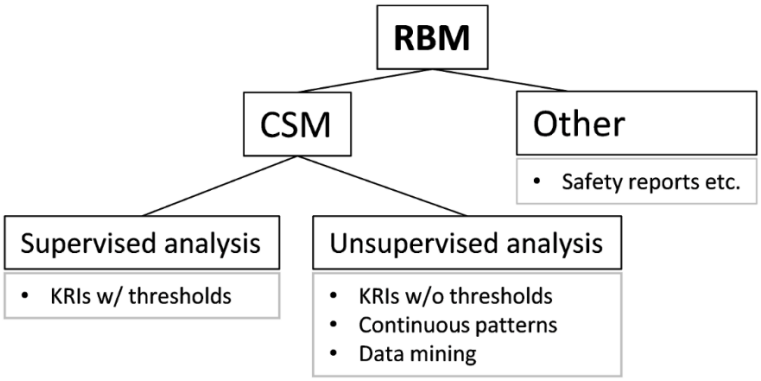


Figure 1. Structure and methods of risk based monitoring (RBM), including central statistical monitoring (CSM) and key risk indicators (KRIs) with (w/) and without (w/o) thresholds.

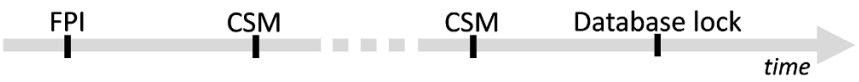


Figure 2. Time flow of a clinical study with respect to central statistical monitoring (CSM) of an ongoing study. FPI; first patient in; meaning the date when the first patient of a study is recruited.

4.2. Supervised analysis

Regarding data quality, supervised analysis can be used to monitor processes (of data collection) and to monitor the data itself. Supervised process monitoring is monitoring of compliance to protocol, meaning how and when procedures are initiated, performed and/or completed in relation to how they ought to be initiated, performed and/or completed. This monitoring is performed to ensure quality in the data collection phase of the process. These procedures could be for example study subject inclusion based on inclusion criteria, study subjects filling out questionnaires, and handling of data queries at the study sites. All this information is collected in software systems and can be extracted as data for analysis.

Supervised data monitoring is performed in a similar manner and concerns for example missing data and the number of data queries (e.g. question about possible data errors). Overall, supervised analysis is based on the use of predefined key risk indicators (KRIs) which are tools created to describe the performance of a site. Example of KRIs are;

- Percent of incorrect inclusions.
- The time from when a study subject is issued a study questionnaire to when it is completed.
- The time from when a study site is sent a query to when it is resolved.
- The amount of missing data or queries.
- The number of reported adverse events.

The selection of which KRIs to investigate is based on the risk that deviation may cause quality loss, as previously recommended (TransCelerate 2014). Thresholds are set for each KRI and visualised as "traffic-light signs" where green corresponds to an interval within a cut-off threshold where a KRI value entails no risk of quality loss, yellow an interval beyond the cut-off where a KRI value entails a possible risk and red an interval beyond a greater cut-off and a probable risk. The obtained result determines if further actions (e.g. directed site visits) are required. The KRI thresholds are set based on various factors; such as the type of study and indication, the patient group and age, since these determine the expected levels of the KRIs. Below follows an invented example of how KRIs can be used to monitor study quality.

An example: a study on a given indication had an expected inclusion rate of 25% with green cut-offs at 20-30%, red cut-offs >40% and <10%, and yellow in between. Supervised analysis was performed a time after the recruitment started. Inclusion percentage was computed for each country and site. All sites in country A had inclusion rates >40% and were flagged red and site B in country B had 35% and was flagged yellow. The rest of the sites had inclusion rates between 20-30% and where flagged green. Further investigation of data from the sites in country A showed that patients here were too generously included into the study. The sites were contacted. It was found that a translation error had led to incorrect inclusion criteria in that country and new criteria was set up. Further investigation of data from site B in country B showed nothing unusual. Additional actions were deemed pending a later RBM.

4.3. Unsupervised analysis

Unsupervised monitoring is statistical analysis of data during the data collection phase without preconception of possible findings. It aims to identify study sites or countries which deviate in data values or data pattern and/or to identify extreme data values or errors. This can be done either with a focus on KRIs, but without predefined thresholds, or by considering all collected data. Simply, this means using statistical methods of either univariate or multivariate analysis (Buyse 1999), to test hypotheses of differences in distribution and to identify outliers in KRI-related data or other data, for example biological measurements. Unsupervised analysis is hence a tool applicable for multi-centre studies with sufficiently large amounts of data for these types of analysis and is therefore typically used for phase III studies. Extreme values can be more easily identified by using unsupervised analysis due to analysis of larger data sets, compared to the traditional approach of local monitoring (as described above). Also, only significantly erroneous data are considered since small errors will potentially be statistically insignificant. This gives the benefit of reducing the amount of work that is not crucial or required. If extreme values or deviant behaviour is found further investigations may be in order.

More complex methods are also available; for example data mining (Venet 2012). In data mining all current data from a study is used to compute an array of p-values and to create a numeric "finger print" for each study site or country. These "finger prints" are then used to identify deviances. Data mining is beyond the scope of this thesis and will not be discussed further here.

5. Methods

5.1. TFS

The current study was conducted in collaboration with TFS (Trial Form Support). TFS is a global mid-size contract research organization with its origin in Lund, Sweden. TFS conducts contracted clinical studies of all phases to aid the life science industry in the pursuit to relieve and treat patients in need. Geographically, TFS has its headquarter in Sweden and have global operations and offices in 21 countries. Their key therapeutic areas include oncology, dermatology, ophthalmology, and cardiovascular disorders, but their experience and competence encompass a wide range of areas.

5.2. SAS

The software used for all statistical analysis and output was SAS Base version 9.4 (TS1M3) with SAS/STAT 14.1 and SAS Enterprise Guide 7.12 by SAS Institute Inc. (Raleigh, NC, USA), except for the output of demographic statistics where Office Excel 2016 by Microsoft Corp. (Redmond, WA, USA) was used.

5.3. The data and descriptive statistics

The data set used in this study was derived as a subsection from a clinical study conducted by TFS. All data, data variables and parameter information were blinded and/or transformed to disable identification of the study, the TFS customer, the study subjects and sites. The data consisting of age of the patients was transformed to not reveal the age group of the patients. The data had been collected at clinical sites in various countries by clinical professionals from recruited patients. The clinical study is conducted according to current regulations and has been approved by relevant global and local ethical committees.

5.4. Choice of variables and flow of CSM analysis

In the current study, data concerning the procedures of adverse events (AE) and serious AE (SAE) were chosen for both supervised analysis and unsupervised KRI-based analysis. Two efficacy biomarkers were chosen for a multivariate unsupervised non-KRI-based analysis. These choices were based on availability of data and determined the detailed structure of the following CSM analysis. Data was extracted at a given time point and prepared for analysis. At an early stage, the question of whether or not all data should be included in the analysis arose. Generally, a certain amount of data is needed for reliable analysis and sites with very short patient time will likely have low amounts of reported events. Therefore, prior to the analysis an inclusion criteria was set. AE and SAE data formed basis for the inclusion criteria, since these variables were the chosen KRIs. A proposed flow of preparation for analysis of AE is given in figure 3. The flow of analysis was designed in order to include only sites where deviations in AE or SAE could lead to risk and therefore to exclude sites with low probability of having any reported AE. Verification of low reporting level among these sites is not of interest, since that is as expected. Identification of high levels of reported AE among these sites could be interesting but they are likely due to random variation. Taken together, these sites were excluded at the given time point (theoretically, to be considered at a later CSM). The inclusion criteria was based on, time to first AE, t_0, which was estimated with 95% probability, as described in the text box below. All sites with a total patient time exceeding the estimated t_0 were included in the CSM analysis. The included data were used for supervised and unsupervised analysis with methods described in detail further on.

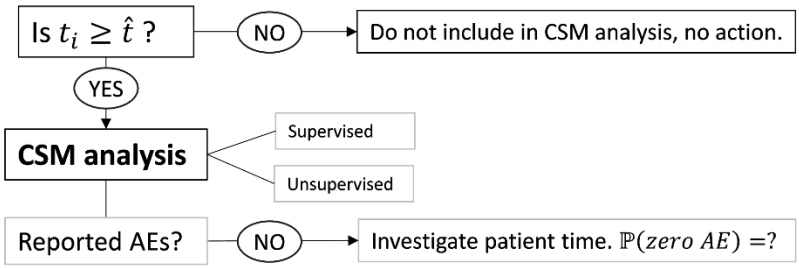


Figure 3. Flow chart of central statistical monitoring (CSM) for adverse events (AE). t_i; total patient time per site at time of CSM, $\hat{t}$; estimated time for at least one event to be probable.

Computation of estimated time to first event (AE or SAE, respectively)

Let $X(t) \in Po(\lambda t)$ where $X(t)$ is a random variable (r.v.) describing the occurrence of (serious) adverse events, λ is the rate of the occurrence and t is time, where $Po(\lambda t)$ is the Poisson distribution with mean λt.

Then, the time between each occurring event is described by r.v. $\tau \in Exp\left(1/\lambda t\right)$, where $Exp\left(1/\lambda t\right)$ is the exponential distribution with mean $1/\lambda t$.

$\hat{\lambda} = MLE(\lambda) = \frac{total\ number\ of\ events}{total\ patient\ time}$

$\mathbb{P}(\tau \leq t_0) = 1 - e^{-\lambda t_0}, where\ t_0\ is\ the\ time\ to\ first\ event.$

Meaning that $\mathbb{P}(X(t_0) = 0) = \mathbb{P}(\tau > t_0)$ and $\mathbb{P}(X(t_0) \neq 0) = \mathbb{P}(\tau \leq t_0)$.

Example: Let $\mathbb{P}(\tau \leq t_0) = 0.95$. Then, the estimated time to first event, $\hat{t}_{0,0.05} = -\ln 0.05/\hat{\lambda}$. Also, $\mathbb{P}(X(t_0) = 0) = 0.05$ and $\mathbb{P}(X(t_0) \neq 0) = 0.95$.

The computed example above concludes that any sites with a total patient time exceeding $\hat{t}_{0,0.05}$ has a less than 5% probability of zero AE, meaning they ought to have at least one AE and are thus included in the CSM analysis. Congruently, $\hat{t}_{0,0.10} = -\ln 0.90/\hat{\lambda}$ is the estimation of time to first event with 90% probability and so on.

Data from all sites with a probability of zero AE of less than 5% were included in the CSM analysis. Data from all sites with a probability of zero SAE of less than 20% were included in SAE rate computations. These levels were chosen to be generous in order to include more data at this investigatory stage. In addition, any included sites that had not reported any AE or SAE were investigated to identify unexpectedly low levels, i.e. large amount of patient time while having no reported AE or SAE, respectively.

5.5. Supervised analysis

The variables AE rate (number of AE/patient time) and SAE rate (number of SAE/patient time) were investigated for threshold-based outcome, where patient time is the sum of all patients' individual time in the study, i.e. the number of days from randomisation to the date when data was extracted for this fictive CSM. In addition, SAE/AE rate was computed for all included sites and countries to provide a

description of AE and SAE reporting behaviours. The thresholds had not been previously defined and therefore various threshold levels were considered. The following ways to set up threshold cut-offs were used:

- Literature-based thresholds; $\pm 15\%$ of the median, as previously described (Oba 2016).
- Intervals based on standard deviation (*std*), such as $\pm\, std$ of the median.
- Intervals based on standard deviation (*std*), such as $\pm\, std$ of the median, on the logarithm of the data.
- Intervals based on median absolute deviation (*mad*), such as $\pm\, mad$ of the median.

The threshold analysis based on median and median absolute deviation (MAD), was defined as visualised in figure 4 and the outcome flagged as GREEN (G; as expected), YELLOW (Y; possible deviation), and RED (R; probable deviation). MAD is the median distance between each data point and the median of the data points. Furthermore, the limits were set by taking into account the right skewness of the data as discussed in section 6.2.

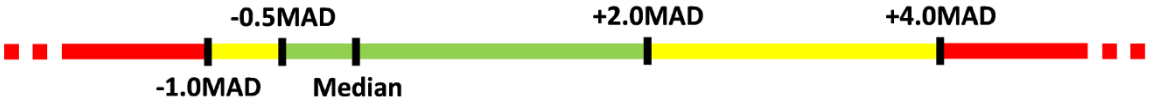


Figure 4. Threshold analysis; limits for the three thresholds; GREEN, YELLOW and RED, as intervals about the median created by $\pm$ numbers of MAD (median absolute deviation).

The normality of data was investigated using descriptive statistics, histograms and qq-plots.

Included sites and countries with zero reported AE were not flagged as above. Instead, unexpectedly low levels of AE were identified by computing the probability to have zero AE. The outcomes were flagged in a similar manner as above, G for >5% probability; Y for 1-5% probability; R for <1% probability. Data was reported only for sites with Y or R flags.

Detailed data of countries and sites with Y and R flags following the supervised analysis were investigated further as deemed appropriate.

5.6. Unsupervised analysis

AE and SAE rates were chosen also for the unsupervised analysis to enable comparison of findings from supervised and unsupervised analysis. The same inclusion criteria as for AE supervised analysis was used. In addition, two efficacy biomarkers were chosen in order to use methods possible on continuous data, namely to perform the proposed multivariate analysis of Mahalanobis distance. The chosen methods have previously been described for use in RBM (Kirkwood 2013, Oba 2016).

5.6.1. Descriptive statistics and identification of outliers

The data was visualised in box plots and scatter plots in order to visually apprehend the data and to notice country and/or site deviant behaviour. AE, SAE and SAE/AE rates as well as the efficacy biomarkers were plotted as a total and by country in boxplots to also identify outliers. AE, SAE and SAE/AE rates were plotted against total patient time per site in scatterplots. Values farther than 1.5*IQR (inter quartile range) from 75-percentile were considered outliers in the box plots.

5.6.2. Chi-square goodness of fit test

The Chi-square goodness of fit test is used to test a difference in observations of frequencies from two (or more) distributions (see text box below for theoretical calculations).

Theory of the Chi-square goodness of fit test

Let n independent observations from a random sample be classified in k number of categories resulting in the values $x_1, \dots, x_k$. Let p_i be the probability that an observation is classified into the i^{th} category. Then, the expected number of observations in each category is $m_i = np_i$ for all i, where

$\sum_{i=1}^{k} p_i = 1$ and $\sum_{i=1}^{k} m_i = n \sum_{i=1}^{k} p_i = n$.

Test null-hypothesis (as relevant in this thesis) H_0: $p_i = {}^1/_k$ for all i with alternative hypothesis H_1: $p_i \neq {}^1/_k$ for some i.

Then, $X^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i}$ is Chi^2 distributed with $k - 1$ degrees of freedom.

5.6.3. One-way analysis of variance

The one-way ANOVA tests difference in mean between independent observations from two or more normal distributions with equal variances where the residuals of the test model are normally distributed (see text box below for theoretical calculations). The normality of the residuals was investigated using histograms and qq-plots. Any highly influential data points were identified using Cook's distance (theoretical details not shown here). The equality variances was tested using a test of homogeneity (Folded T test; theoretical details not shown here).

Theory of the one-way ANOVA (analysis of variance)

Let $y_{A1}, \dots, y_{An}$, $y_{B1}, \dots, y_{Bn}$ and $y_{C1}, \dots, y_{Cn}$ be independent observations of the random variables $Y_A \in \mathcal{N}(\mu_A, \sigma^2)$, $Y_B \in \mathcal{N}(\mu_B, \sigma^2)$ and $Y_C \in \mathcal{N}(\mu_C, \sigma^2)$ with unknown μ_A, μ_B, μ_C and σ^2. Estimations of the treatment means are $\hat{\mu}_A = \bar{y}_A, \hat{\mu}_B = \bar{y}_B, \hat{\mu}_C = \bar{y}_C$ and of the grand mean is $\hat{\mu} = \bar{y}$.

Test null-hypothesis H_0: $\mu_A = \mu_B = \mu_C$ with alternative hypothesis H_1: $\mu_i \neq \mu_j$, for some $i \neq j$.

The ANOVA model is built so that the data deviation from grand average equals the treatment deviation from grand average plus a residual.

Estimation of σ^2 between treatments is $s_T^2 = \frac{\sum_{j=A}^{C} (\bar{y}_j - \bar{y})^2}{3-1}$.

Estimation of σ^2 within treatments (of residuals) is $s_R^2 = \frac{\sum_{j=A}^{C} \sum_{i=1}^{n} (y_{j,i} - \bar{y}_j)^2}{3n-3}$.

Then, $F_0 = ns_T^2 / s_R^2 \in F(3-1, 3(n-1))$.

Post hoc analysis was performed on any statistically significant result in the ANOVA analysis; meaning for example if an ANOVA showed statistically significant difference in means between countries the post hoc analysis was used to identify which country that was significantly different from the rest. The

choice of post hoc analysis; here unpaired t-test or Wilcoxon ranked-sum-test, depended on the data distributions as described below.

5.6.4. Unpaired t-test

The unpaired t-test was used to test difference in means between independent observations from two normal distributions with equal variance (see text box below for theoretical calculations). The normality of data was investigated using descriptive statistics, histograms and qq-plots and the equality of variances was tested using Levene's test (detailed descriptions not given here).

Theory of the unpaired t-test

Let $y_1, \dots, y_{n_Y}$ *and* $x_1, \dots, x_{n_X}$ be independent observations of the random variables $Y \in \mathcal{N}(\mu_Y, \sigma^2)$ and $X \in \mathcal{N}(\mu_X, \sigma^2)$ with unknown μ_Y, μ_X, and σ^2. Estimations of the unknowns are $\hat{\mu}_Y, \hat{\mu}_X$, and s^2, respectively,

where $\hat{\mu}_Y = \bar{y}, \hat{\mu}_X = \bar{x}$, and $s^2 = \frac{\sum_{i=1}^{n_Y}(y_i-\bar{y})^2+\sum_{i=1}^{n_X}(x_i-\bar{x})^2}{n_Y+n_X-2}$.

Test null-hypothesis H_0: $\mu_Y = \mu_X$ with alternative hypothesis H_1: $\mu_Y \neq \mu_X$.

Then, $t_0 = \frac{(\bar{y}-\bar{x})-0}{s\sqrt{\frac{1}{n_Y}+\frac{1}{n_X}}} \in t(n_Y + n_X - 2)$.

5.6.5. Wilcoxon ranked-sum test

The Wilcoxon ranked-sum test (Wilcoxon test), also called the Mann Whitney U test, is an alternative to the unpaired t-test. It is used to test difference in rank between observations from two distributions without assuming normal distribution (see text box below for theoretical calculations).

Theory of the Wilcoxon ranked-sum test

Let $y_1, \dots, y_n$ *and* $x_1, \dots, x_n$ be independent observations of the random variables $Y \in A$ and $X \in B$ where A and B are unknown distributions with equal variances.

Test null-hypothesis H_0: $A = B$ with alternative hypothesis H_1: $A \neq B$.

Let there be a list of all observations ordered in ascending order. Then, rank is a number from 1 to 2n given in order to the items of the list with 1 assigned to the smallest observation and so on.

The sum of ranks for the observations in Y is $w_Y = \sum_{i=1}^{n} the\ rank\ of\ each\ y_i$, where $w_Y \in W_Y$, and equivalently for X.

Then, $\alpha = 2\mathbb{P}(W_Y \geq w_Y)$ for upper tail or $= 2\mathbb{P}(W_Y \leq w_Y)$ for lower tail, or equivalently for X.

5.6.6. Mahalanobis distance

Mahalanobis distance is an entity that describes the number of standard deviations between a data point in a multivariate distribution and the mean of the distribution taking into account the different variances of the variables (De Maesschalck 2000). All variables are assumed to be normally distributed and if their variances are equal then Mahalanobis distance is reduced to the Euclidian distance. If computed

for a univariate standard normal distribution, $\mathcal{N}(0,1)$, the distance is reduced to the (positive) z-score. Detailed information on the calculation of Mahalanobis distance is given in the text box below. Since the Mahalanobis distance is a measure of distance from the mean only outliers away from zero (and not towards zero) are considered relevant.

Theory of Mahalanobis distribution

Let $x_{i,j}$ be the i[th] independently drawn observation *(i=1,…,n)* on the j[th] random variable *(j=1,…,p)* with multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Then, the sample mean vector, $\bar{\boldsymbol{x}}$, is a vector whose j[th] element is the average value of the n observations of the j[th] variable:

$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}$, the mean vector is given by: $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_j \\ \vdots \\ \bar{x}_p \end{bmatrix}$, and the sample covariance matrix is an $p \times p$ matrix, $\boldsymbol{Q}$, given by: $\boldsymbol{Q} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_{i\cdot} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{i\cdot} - \bar{\boldsymbol{x}})^T$, with each entry: $\boldsymbol{q_{j,k}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{i,j} - \bar{x}_j\right)\left(x_{i,k} - \bar{x}_k\right)^T$

Given an observation $\boldsymbol{x}_i$ from the distribution above, the squared Mahalanobis distance (d) from $\boldsymbol{x}_i$ to $\boldsymbol{\mu}$ is the number of standard deviations between the two. This is given by the formula:

$$d = \sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})}$$

, where

$$\Sigma = E[(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T] = \begin{bmatrix} E[(x_1 - \mu_1)(x_1 - \mu_1)] & \cdots & E[(x_1 - \mu_1)(x_n - \mu_n)] \\ \vdots & \ddots & \vdots \\ E[(x_n - \mu_n)(x_1 - \mu_1)] & \cdots & E[(x_n - \mu_n)(x_n - \mu_n)] \end{bmatrix}$$

which is estimated by $\boldsymbol{Q}$ and where $\boldsymbol{\mu}$ is estimated by $\bar{\boldsymbol{x}}$.

5.6.7. Alpha-level

An alpha-level of 5% was considered statistically significant for all tests. The alpha-level is the risk of rejecting the null-hypothesis when indeed the null-hypothesis is true. The null-hypothesis in all performed tests was no difference in mean, rank, or distance.

6. Results

6.1. The data and descriptive statistics

The data set used in this study was a subsection of the data of a clinical study conducted by TFS. At the time of data extraction the patient demographic description of this subsection was as described in total and by country in table 1 (demography data by site is not shown). The sectioned data had been collected from 75 sites in 11 countries with 1-8 sites per country. The patient group consisted of more women than men. Age and sex proportion varied across the recruiting countries; age was significantly lower in Country 04 and 07 compared to the rest (t-test; data not shown) and the proportion female patients was significantly high in Country 08 while not in the other countries. The data was collected during approximately the same time period in all countries except for one country which was opened later and where the first randomisation took place more than 6 months later than the rest. From the above data,

51 sites in ten countries were included in the analysis and, from these, a total of 685 AE and 43 SAE had been reported over 97548 patient days.

Table 1. Demographic details of the study group in total and by country.

	Patient time, days (%)	**Age, mean (std)**	**Female, %**	**Chi^2 (df, p)**
Total	103352 (100)	4.32 (0.11)	56.1	10.1 (1, 0.002)
Country:				
-01	240 (0.2)	4.26 (0.14)	25.0	1.00 (1, 0.32)
01	11235 (10.9)	4.33 (0.10)	57.1	1.29 (1, 0.26)
02	19765 (19.1)	4.34 (0.12)	52.9	0.29 (1, 0.59)
03	4251 (4.1)	4.36 (0.09)	60.0	0.80 (1, 0.37)
04	14675 (13.2)	4.28 (0.11)	60.0	2.20 (1, 0.14)
05	2794 (2.7)	4.34 (0.15)	58.3	0.33 (1, 0.56)
06	8900 (8.6)	4.34 (0.11)	38.1	2.38 (1, 0.12)
07	881 (0.9)	4.22 (0.12)	33.3	1.00 (1, 0.32)
08	11379 (11.0)	4.30 (0.11)	70.2	9.28 (1, 0.002)
09	18222 (17.6)	4.30 (0.12)	51.3	0.05 (1, 0.82)
10	9212 (8.9)	4.34 (0.09)	64.9	3.27 (1, 0.07)
11	1798 (1.7)	4.34 (0.11)	85.7	3.57 (1, 0.06)
F (df_1, df_2, p) =		4.43 (11, 711, <0.0001)		

Patient time: the sum of days in study for each patient. Age: transformed actual age. df: degrees of freedom. p: p-value. The proportion of females in each country was tested with Chi^2 test and any difference in age between the countries was tested with ANOVA (F statistic).

6.2. Supervised analysis – Adverse events

The time to first event, AE and SAE respectively, was estimated as described in the methods section to establish which sites to include in the CSM analysis (table 2, see Appendix 3.1 for the SAS code).

Based on the inclusion criteria, 24 sites (including one country) were excluded from the CSM analysis (appendix 2.1). AE and SAE/AE rates were investigated on data from the remaining 51 sites in ten countries. SAE rate was investigated on data from five individual sites and on summarised data from eight countries.

Table 2. Estimation of time to first event, $\hat{t}_{0,\alpha}$ (days), via estimation of the rate of events, $\hat{\lambda}$, for adverse events (AE) and serious adverse events (SAE), respectively.

	AE	**SAE**
$\hat{\lambda} =$	0.006889 ...	0.0004354 ...
$\hat{t}_{0,0.05} =$	434.9	6880
$\hat{t}_{0,0.20} =$	–	3696

The starting point was to use literature-based thresholds, such as $\pm 15\%$ of the median, as previously described (Oba 2016). However, this resulted in very small intervals about the median with signals for further investigation of nearly all sites and counties (data not shown). The next idea was to create intervals based on standard deviation (std), such as $\pm 1.5\ std$ of the median. This resulted in a bulk of sites and countries not signalled for further investigation (data not shown) but an assumption for this threshold

analysis was to have symmetrically distributed data. By nature, this is of course not the case for AE and SAE rates which can be assumed to be Poisson distributed or seen as very right-skewed normal. Therefore, threshold analysis was additionally performed on transformed; logarithmic, data. Transformation made AE data fairly symmetric since a rather large number of AE were reported (figure 5). However, for SAE the transformation did not lead to a symmetric distribution and led to a high proportion of missing data since many sites had zero SAE. Still, it may be possible that by combining threshold analysis on both non-logarithmic and logarithmic data sufficient outputs would be produced. Still, the analysis would have to be adjusted for each parameter and was not considered general enough. The chosen and reported threshold analysis was based on median (m) and median absolute deviation (mad), defined as visualised in figure 4 and the outcome flagged as GREEN (G; as expected), YELLOW (Y; possible deviation), and RED (R; probable deviation). This allowed for more efficient investigation of non-normal data since mad is the median distance between each data point and the median of the data points. Furthermore, the limits were set by taking into account the right skewness of the data; values $> median(m) - 0.5\ mad$ and $< m + 2.0\ mad$ flagged G; values $< m - 1.0\ mad$ and $> m + 4.0\ mad$ flagged R, and values in between flagged Y.

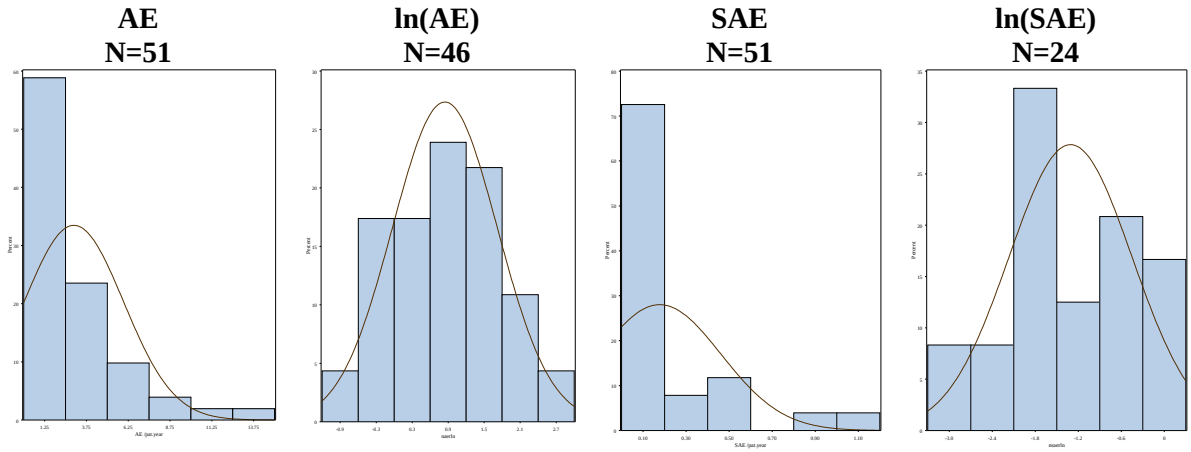


Figure 5. CSM – Supervised analysis; histograms of AE and SAE data with no transformation and logarithm-transformed (ln), respectively. Axis are not described in detail since considered unimportant and approxiamtions are possible.

Country-wise, the supervised CSM analysis revealed a probable data deviation in Country 10 and possible data deviations in countries 01, 04, and 07 (table 3). Country 10 had a rate higher (R) than average and the latter countries had lower (Y). Country 04 had also a low SAE rate (R) compared to average while all other countries had SAE rates as expected. At site-level, nearly all sites in Country 10 had individually high levels (Y or R) compared to average (appendix 2.2). Apart from those in Country 10, another four sites had high AE rates (Y) and thirteen had low AE rates (R or Y). One site had high SAE rate (R) compared to average and two sites had low (Y).

No countries were flagged for low probabilities of zero AE or SAE, respectively, and no sites were flagged for low probability of zero SAE (appendix 2.2). Of the sites with zero AE, three were flagged (R); C01-S04 with $\mathbb{P}(zero\ AE) = 0.48\%$, C04-S02 and C04-S05 both with $\mathbb{P}(zero\ AE) < 0.01\%$; and two were flagged (Y); C03-S08 with $\mathbb{P}(zero\ AE) = 1.1\%$ and C07-S01 with $\mathbb{P}(zero\ AE) = 1.7\%$.

A summary table of all findings from KRI-based analysis on AE and SAE is given in appendix 2.5.

Table 3. CSM – Supervised analysis; adverse events summary; total rates and rates by country.

	Patient time (days)	**AE**	**SAE**	**AE /pat.year**	**SAE /pat.year**	**SAE/AE**
Total	97548	685	43	2.56	0.161	6%
Country 01	9880	19	2	0.70-Y	0.074-Y	11%
Country 02	18851	139	11	2.69	0.203	8%
Country 03	4077	33	3	2.96	0.258	9%
Country 04	14360	25	1	0.64-Y	0.025-R	4%
Country 05	2112	22	2	3.80		9%
Country 06	8900	75	3	3.08	0.123	4%
Country 07	588	0	0	1.74%-Y		0%
Country 08	10615	46	4	1.58	0.16	9%
Country 09	17489	96	12	2.00	0.241	13%
Country 10	8878	210	3	8.64-R	0.119	1%
Country 11	1798	20	2	4.06		10%

AE: number of AE, SAE: number of SAE, /pat.year: per patient year. Probability flags; Y (YELLOW); Prob(no AE or SAE) 1-5%, R (RED); Prob(no AE or SAE) <1%. Rate flags; Y (YELLOW): Rate value in intervals -1.0MAD to -0.5MAD or +2MAD to +4MAD from median. R (RED): Rate <1.0MAD or >4MAD from median. GREEN flags are not indicated. SAE rates were investigated only if the total patient time was sufficiently long (Prob(no SAE) <20).

6.3. Unsupervised analysis – Adverse events

The boxplot of AE rates showed that Country 10 may have high levels compared to total and Countries 01, 04, and 07 may have low (figure 6). Congruently, one-way ANOVA on AE rates was significant depending on country ($F_{10,40}$=8.91; p<0.0001, r^2=0.69). The residuals from fitting the one-way ANOVA model was found normally distributed and no highly influential observations as seen by the Cook's distance (figure 7). Variances were significantly different between countries but the absolute differences were considered negligible (data not shown). However, the post hoc analysis (Wilcoxon) between each country and the remaining countries together showed that Country 01 (Z=-1.84; p=0.03), Country 04 (Z=-2.60; p=0.005) and Country 10 (Z=4.08; p<0.0001) differed from the rest while the remaining countries did not have deviant AE rates (data not shown). In the total, two sites were outliers; Country 10 Site 02 (C10-S02) and C10-S07, but they were not outliers within Country 10 alone. In addition, sites C01-S08 and C02-S07 were outliers within their respective countries, but not in the total. The scatter plot of AE rates versus total patient time showed that sites C01-S03, C04-S01, C04-S02, C04-S05, C04-S06, and C09-S06 may have low rates and that at least three sites in Country 10 may have high rates (figure 8).

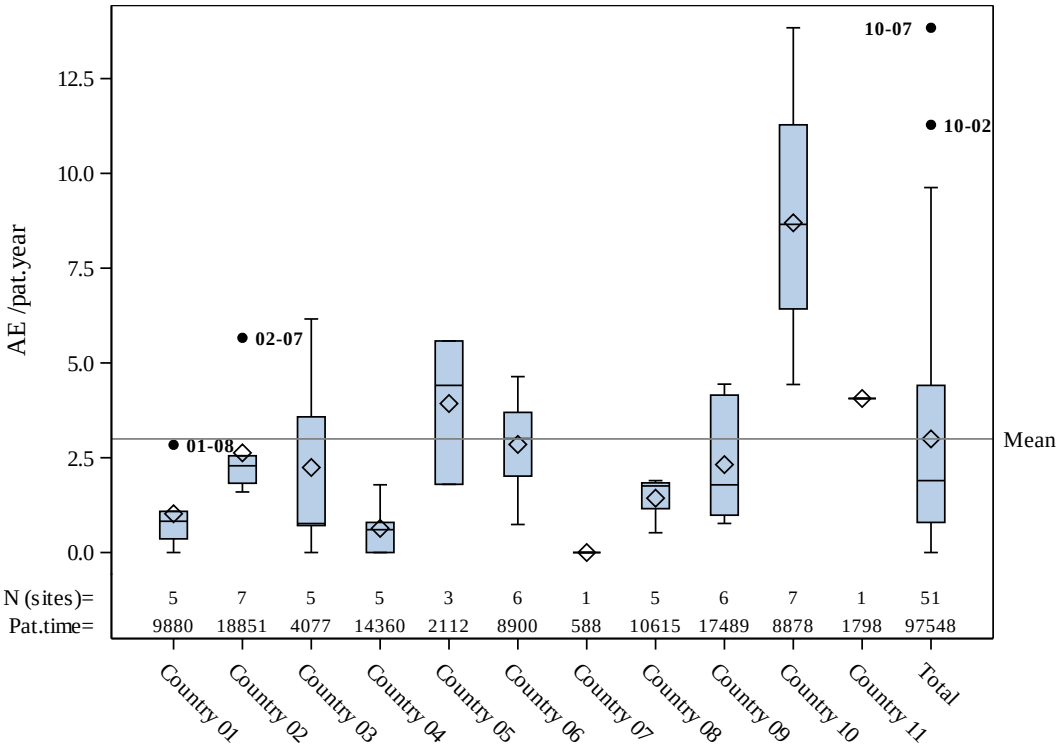


Figure 7. CSM – Unsupervised analysis; boxplot of AE rates, by country and total. Individual site values are marked (country number – site number) if outliers; i.e. values farther than 1.5*IQR (inter quartile range) from 75-percentile. Refline is mean of total.

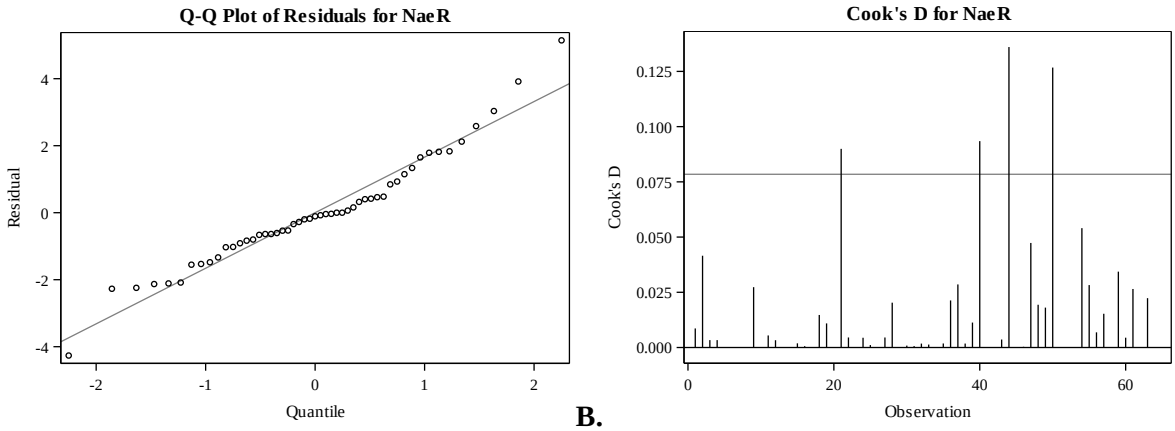


Figure 7. CSM –Unsupervised analysis; diagnostics of one-way analysis of variance of AE rates (NaeR) including QQplot of residuals (A) and Cook's distance (B) showing no highly influential data points.

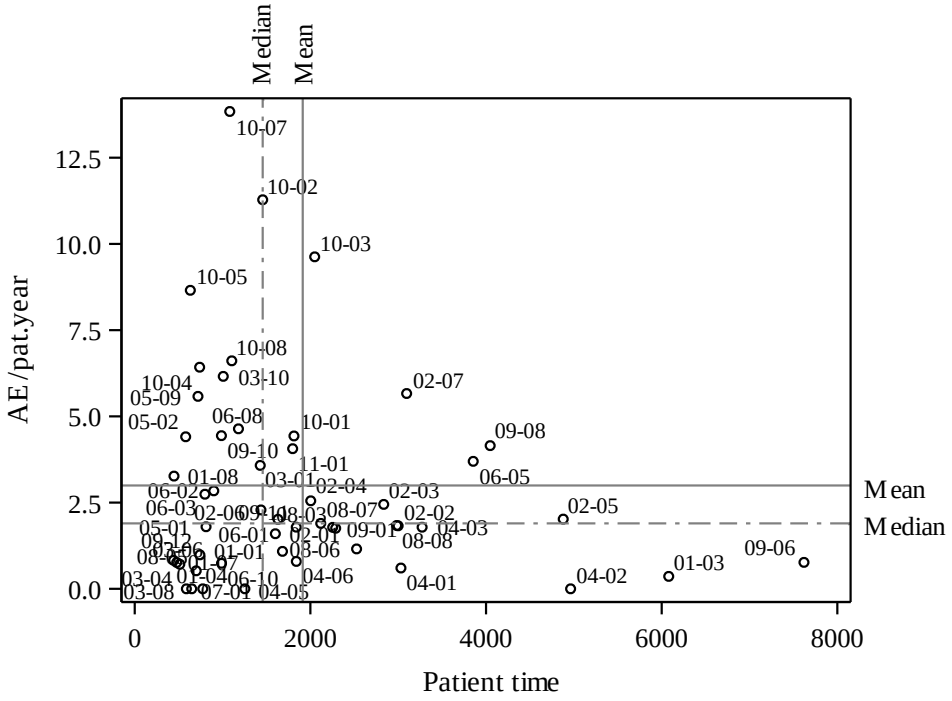


Figure 8. CSM – Unsupervised analysis; scatter plot of AE rates versus patient time (days). /pat.year: per patient year. Reflines are mean and median of AE/pat.year and Patient time, respectively.

The boxplot of SAE rates showed large variation between countries with no obvious trends by individual countries (figure 9). This was confirmed by the one-way ANOVA on SAE rates which did not show significant difference depending on country ($F_{10,40}$=0.50; p=0.8817). The residuals from fitting the one-way ANOVA model was found to be questionably normally distributed and no highly influential observations as seen by the Cook's distance (figure 10). Variances were not significantly different between countries (data not shown). In the total, sites C03-S10, C06-S02, C09-S08, and C10-05 were outliers with high rates and all but C09-S08 were outliers in their countries, respectively, as well. Within each country, two additional outliers with high rates (C01-01 and C04-S03) were found and one with low rate (C08-S05). The scatter plot of SAE rate versus total patient time showed that sites C02-S02, C04-S01, C04-S02, C01-S03, and C09-S06 may have low rates and that sites C03-S10, C06-S02, C09-S08, and C10-S05 may have high rates (figure 11).

Similar to that of SAE rates, the boxplot of SAE/AE showed large variation and no obvious trends by countries (appendix 2.3), confirmed by the one-way ANOVA test on SAE/AE which did not show significant differences depending on country ($F_{10,40}$=0.81; p=0.6239; residual diagnostics not shown). In the total, sites C05-S01, C06-S02, and C09-S12 were outliers with high SAE/AE of which the latter two were outliers within their countries, respectively, as well. In addition, sites C03-S10, C04-S03, and C10-S05 were outliers with high SAE/AE and C08-S05 with low SAE/AE, within their countries, respectively. Scatter plot of SAE/AE showed that sites C02-S02, C04-S01, and C04-S02 may have low SAE/AE rate and that site C09-S12 may have high (appendix 2.4).

A summary table of all findings from KRI-based analysis on AE and SAE is given in appendix 2.5.

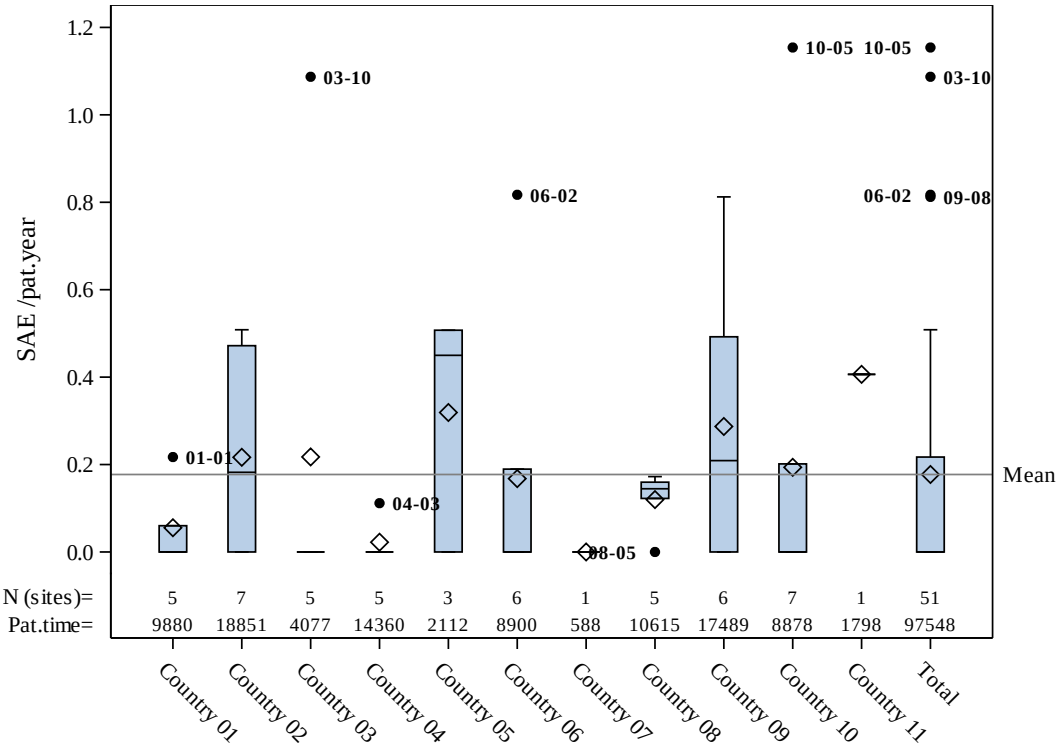


Figure 9. CSM – unsupervised; boxplot of SAE rates by country and total. Individual site values are marked (country number – site number) if outliers; i.e. values farther than 1.5*IQR (inter quartile range) from 75-percentile. /pat.year: per patient year. Refline is mean of total.

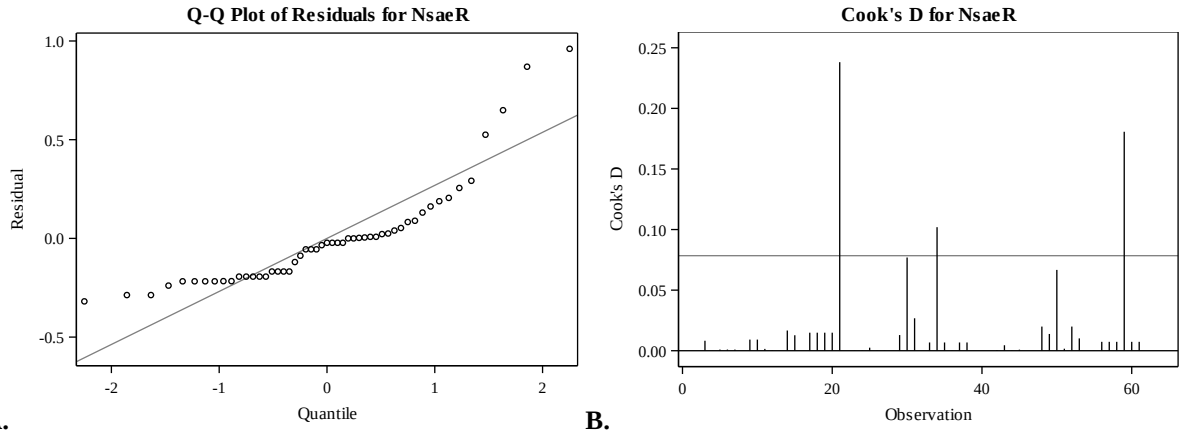


Figure 10. CSM –Unsupervised analysis; diagnostics of one-way analysis of variance of SAE rates (NsaeR) including QQplot of residuals (A) and Cook's distance (B) showed no highly influential data points.

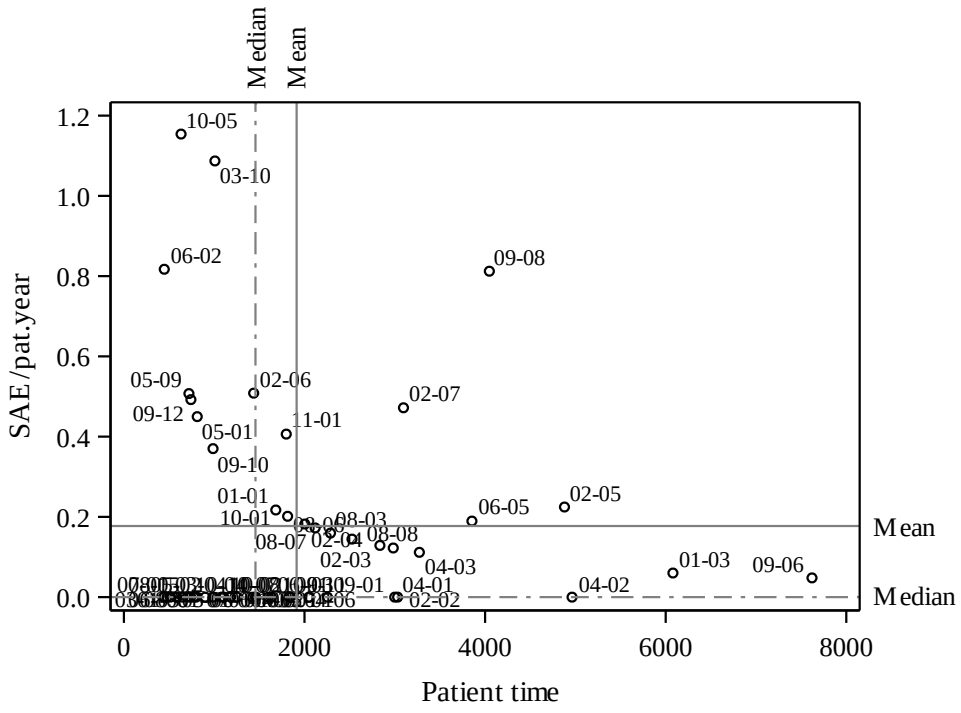


Figure 11. CSM – unsupervised; scatter plot of SAE rates versus patient time (days). /pat.year: per patient year. Reflines are mean and median of SAE/pat.year and Patient time, respectively.

6.4. Unsupervised analysis – Efficacy biomarker

The box plots of efficacy biomarker 1 (eff1) and 2 (eff2) were alike in distributions (eff1; figure 12A, eff2; appendix 2.6) and closely correlated (data not shown). Countries 01 and 06 seemed to have low levels and countries 03, 07, 08, and 11 seemed to have high levels compared to total. One-way ANOVA tests confirmed a difference in levels depending on country (eff1; $F_{10,430}$=2.74; p=0.0028; r^2=0.06, eff2; $F_{10,430}$=2.81; p=0.0022; r^2=0.06; residual diagnostics not shown). The post hoc analysis (Wilcoxon) showed that Country 01 had significantly lower levels of both biomarkers (eff1; 27.8 vs. 32.1; Z=-3.51; p=0.002, eff2; 57.5 vs. 62.1; Z=-3.51; p=0.002), Country 04 had higher levels (eff1; 33.5 vs. 31.3; Z=1.73; p=0.042, eff2; 63.5 vs. 61.3; Z=1.73; p=0.042) as did Country 11 (eff1; 36.9 vs. 31.5; Z=1.67; p=0.048, eff2; 66.9 vs. 61.5; Z=1.67; p=0.048), compared to the remaining countries, respectively. No other countries had significantly different levels (data not shown). Further tests on these countries, respectively, showed no significant differences depending on site (data not shown). One-way ANOVA tests showed significant differences in eff1 and eff2 depending on site in the whole data set (eff1; $F_{50,390}$=1.59; p=0.0087; r^2=0.170, eff2; $F_{50,390}$=1.60; p=0.0083; r^2=0.170). Site C03-S06 was outlier with high eff1 and eff2 levels, in the total groups as well as in Country 03. It was confirmed to have significantly higher eff1 and eff2 levels compared to the rest (eff1; 42.0 vs. 32.5; Z=2.66; p=0.0039, eff2; 72.0 vs. 61.4; Z=2.66; p=0.0039). No other sites were tested.

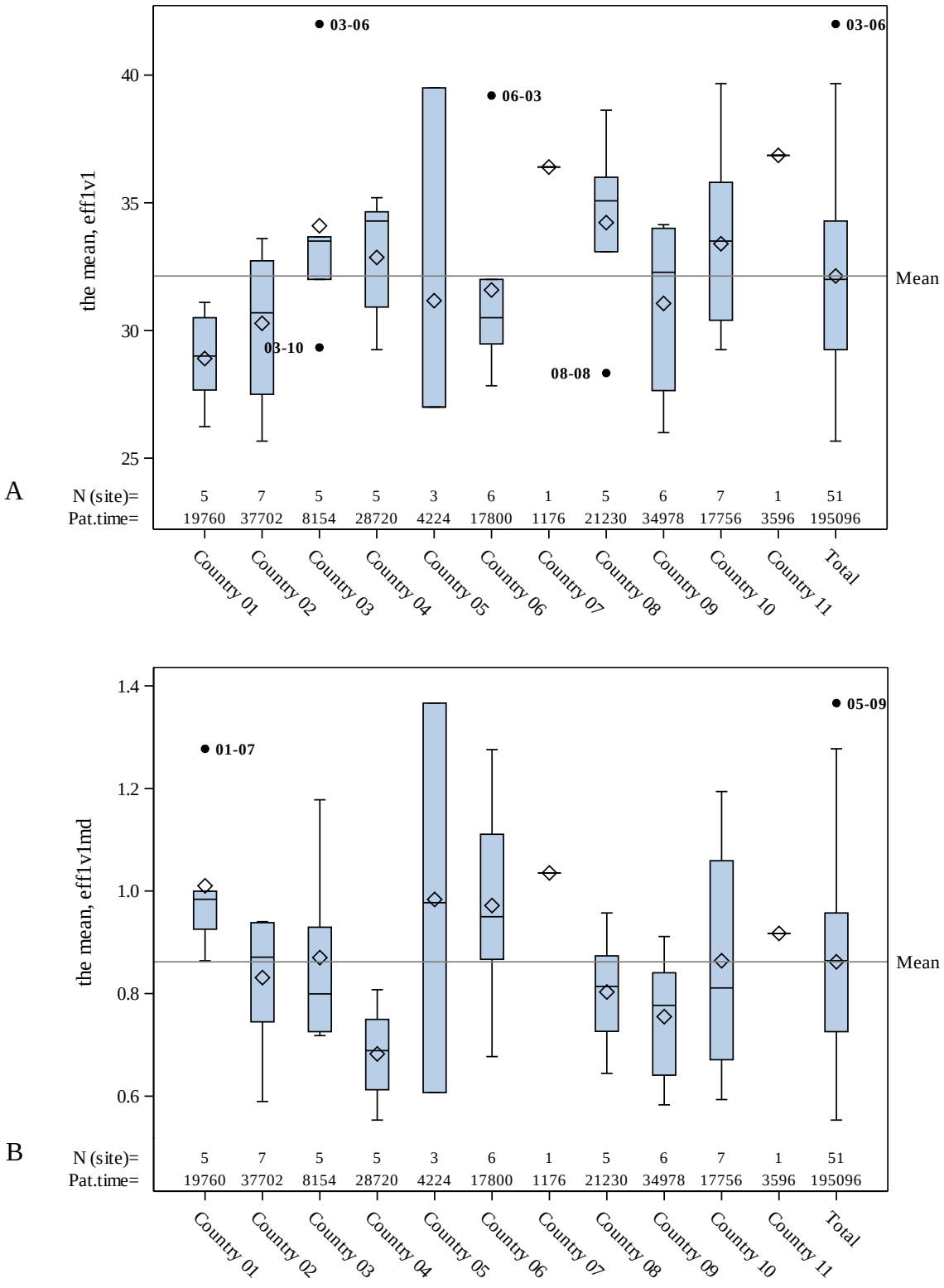


Figure 12. CSM – Unsupervised analysis; efficacy biomarker 1 (A; eff1v1) and Mahalanobis distance of the same (B; eff1v1md) across countries; means for each site. Number (N) of patients and sites are given for each country, respectively. Individual site values are marked (country number – site number) if outliers; i.e. values farther than 1.5*IQR (inter quartile range) from 75-percentile. Refline is mean of total.

Similarly to above, the box plots of the univariate Mahalanobis distance for eff1 (eff1md) and eff2 (eff2md), respectively, were alike in distribution (eff1md; figure 8B, eff2md; appendix 2.7). Larger Mahalanobis distance than total (which is the relevant direction to study for Mahalanobis distance) was seen for Country 01, 06, 07, and 11. However, testing could not identify any significant differences depending on country or site for eff1md or eff2md (data not shown). The box plots of the multivariate Mahalanobis distance for eff1 and eff2 showed that Country 01 and possible 06 and 07 had large distance compared to total (figure 13). However, as for the univariate analysis, testing did not confirm any significant differences (data not shown). Still, site C05-S09 was an outlier in the total of the univariate Mahalanobis distances and site C10-S04 in the multivariate.

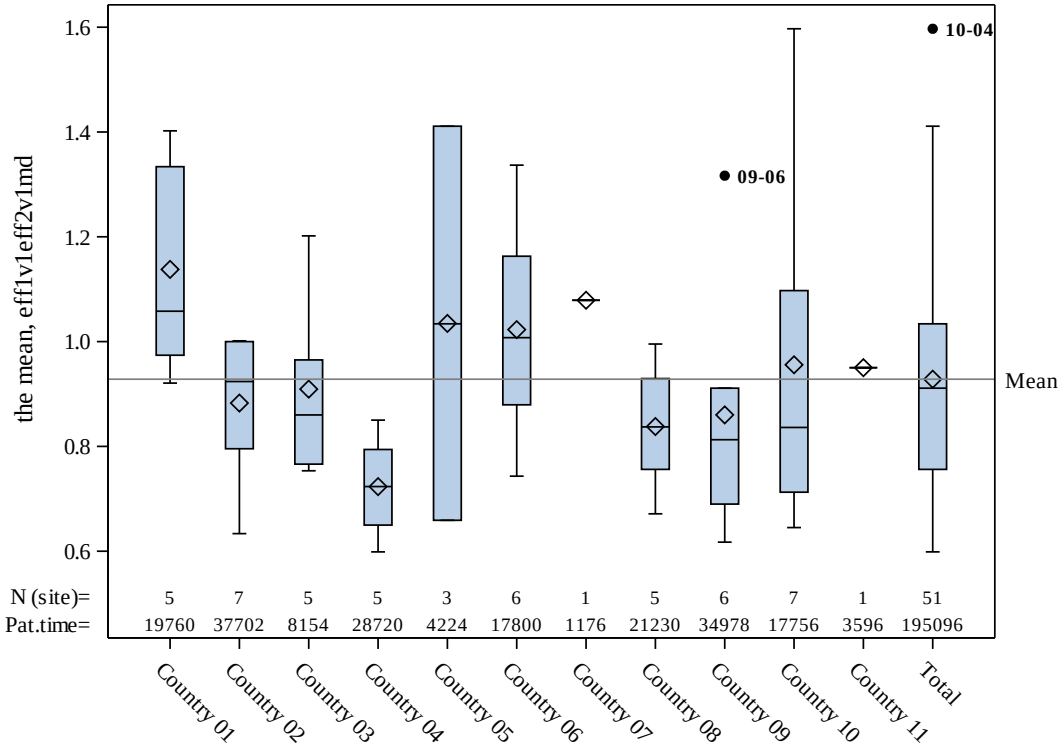


Figure 13. CSM – Unsupervised analysis; Mahalanobis distance for efficacy biomarker 1 and 2 (multivariate analysis) across countries; means for each site. Number (N) of patients and sites are given for each country, respectively. Individual site values are marked (country number – site number) if outliers; i.e. values farther than 1.5*IQR (inter quartile range) from 75-percentile. Refline is mean of total.

7. Discussion and Conclusions

7.1. Data findings

7.1.1. Adverse events

A summary of all countries and sites with deviant data from KRI-based analysis of adverse events (both AE and SAE, and both supervised and unsupervised analysis) is given in appendix 2.5. From supervised analysis, twenty-seven sites in ten countries were flagged as having deviant data in either AE or SAE and of these, three countries as a whole were flagged. This is a rather large proportion of the data set; nearly 50% of the sites were flagged (36% of the sectioned data). This implies that the threshold cut-offs were probably too generous. Especially since the cut-offs were based on the data itself and not prior

studies. Nine sites were flagged red (and no countries) which may constitute a more reasonable level of deviations. Seven of these sites were identified also in unsupervised analysis which, in total, resulted in twenty sites in eight countries being flagged as having deviant data in either AE or SAE. Also, the same three countries as in supervised analysis were identified in unsupervised analysis. However, since unsupervised analysis on SAE was done using the AE-based inclusion criteria (which for example supervised analysis on SAE was not) many observed outliers were due to deviant SAE rates in sites with short patient time. Actually, only one site in the data set had sufficient patient time for a probability of zero SAE of less than 5%. Furthermore, outliers within each country that were not outliers in the total group may likely be disregarded. Guided by this reasoning, the unsupervised analysis resulted in that seven sites only were flagged as deviant in unsupervised analysis. As expected, most findings in supervised and unsupervised analysis overlapped and the combined findings resulted in eleven sites in five countries being flagged which was equivalent to 19% of the sites that passed the inclusion criteria and 15% of the sites in the sectioned data. The final findings are summarized in table 4.

Table 4. Final summary of selected KRI-based (supervised and unsupervised analysis) results. All given sites were flagged in supervised analysis. Squares mark findings in also unsupervised analysis.

		Patient time (days)	**AE**	**SAE**	**AE /pat.year**	**SAE /pat.year**	**Comments**
Country 01		9880	19	2	0.70-Y		Overall low levels in Country 01 with no significant difference between sites.
Country 01	Site 03	6080	6	1	0.36-R	0.060-Y	
Country 01	Site 04	775	0		0.48%-R		
Country 04		14360	25	1	0.64-Y	0.025-R	Overall low levels in Country 04 with no significant difference between sites.
Country 04	Site 01	3031	5	0	0.60-R		
Country 04	Site 02	4963	0		0.00%-R		
Country 04	Site 05	1254	0		0.02%-R		
Country 08	Site 05	700	1	0	0.52-R		Low levels at site C08-S05.
Country 09	Site 06	7620	16	1	0.77-Y	0.048-Y	Low levels at site C09-S06.
Country 10		8878	210	3	8.64-Y		Overall high levels in Country 10 with no significant difference between sites.
Country 10	Site 02	1457	45	0	11.28-R		
Country 10	Site 03	2049	54	0	9.63-R		
Country 10	Site 05	633	15	2	8.66-R		
Country 10	Site 07	1082	41	0	13.84-R		

AE: number of AEs, SAE: number of SAEs, /pat.year: per patient year. Y; yellow flag, R; red flag.

Flags due to low levels signals for further investigations at each site, since more information is unlikely found in the data. However, situations with flags due to high levels may benefit from further data investigations. These findings may be caused by “harmless” errors such as technical duplications or reporting traditions which may be revealed by investigating the data more closely. In fact, sites in Country 10 seem to be pertinent in reporting one adverse event every time a patient experience it leading to several reported events where other sites may report this as one event only.

How well the findings presented here correspond to previous findings is unclear since no previous publications reporting the outcome of KRI-based supervised analysis of adverse events have been found. Supervised and unsupervised analysis gave largely the same result, meaning the flagged sites and countries of the two overlapped. On one hand, this could indicate that the results were reliable. On the other hand, one could argue that the two types of analysis used similar methods (although different tests) since the supervised analysis method was based on the current data and not previously validated.

7.1.2. Efficacy biomarkers

Unsupervised analysis of the efficacy biomarkers resulted in identification of three potentially deviant sites; a different one for each type of analysis (raw data, univariate and multivariate Mahalanobis distance, respectively). However, the absolute differences were small and may be negligible depending on the nature of the efficacy biomarkers. Therefore, it is difficult to speculate if these variations are of importance. The two biomarkers in this study were highly correlated. This correlation may be the reason as to why multivariate analysis did not result in further information compared to univariate. Overall, the data of the two efficacy biomarkers seemed to have no evident risk of quality loss. Still, improved methods for unsupervised analysis may have revealed other findings.

Outliers were investigated to find extreme patients or sites or countries and potential differences between sites and countries, respectively, were tested. Another interesting aspect in this data would have been to investigate rounding errors by considering digit preference, as previously described (Al-Marzouki 2005).

7.2. Method findings

7.2.1. Supervised analysis

Performing KRI-based CSM with pre-set thresholds may seem to be a straightforward process. However, a major difficulty lies in setting up the method which needs to be pre-defined, programmed, tested, and validated which has previously been noted (Buyse 2014). Which methods to use for the set-up depends firstly on which KRI to investigate, how these are defined and which deviations of these are important to identify. One aspect is to set the threshold limits. When analysing AE and SAE, both high and low deviations are of interest. High levels of AE are normally not a problem but may pose problems for reliable efficacy evaluation (due to for example extreme patients). High levels of SAE is a concern since this may signal an unexpectedly unsafe patient situation or a misunderstanding of the concept SAE. Low levels of AE or SAE are of crucial importance to identify since this may be due to failure to report these events and thus cause safety and quality problems. While both directions are important, safety is top priority in clinical studies and therefore low levels of adverse events are of especial importance to identify. This is challenging since the reporting of adverse events is a counting value starting from zero and a certain minimum exposure time is needed before AE or SAE can be expected. Hence, early in a study too low reported levels may be hidden by the fact that low levels are likely. In the current study several thresholds were considered in order to find appropriate ones and the use of combined non-log and log may be valuable but should perhaps not be considered for data with high proportion of zero valued data. The use of median absolute deviation with arbitrary cut-off limits, as currently used, may be a reasonable solution. However, by using statistics as basis for the threshold some sites will always be flagged as deviant in the given data. By using a completely arbitrary basis for the thresholds one may gain more relevant results. Still, it does not solve the problem with having many zero-valued data points and the choice of cut-off limits is very sensitive to this problem on the left side of the distribution. Here, the yellow flags may have been too generous while the red flags may have been reasonably defined. The cut-off limits need to be further refined and verified by evaluating what a detailed site investigation actually brings after the site being signalled as possibly deviant.

Another challenging aspect in KRI-based CSM is the inclusion criteria. Early in a study some sites may have only a handful of recruited patients and limited amounts of data. They may have no reported AE and this may be normal depending on total patient time, indication etcetera. However, this poses a dilemma as to how to deal with them statistically. This dilemma goes hand-in-hand with the one described above and stems from the facts that the levels of adverse events (as many other KRIs) is a counting value. In the current study, an inclusion criteria based on exposure time set from the estimated probability to have an event was used. This seems reasonable but separate inclusion criteria would be

required for each KRI to investigate. SAE are less frequent than AE. Hence, the problems described above are greater for SAE than for AE and greater efforts are needed to set up the thresholds for SAE. It may be questionable if reliable deviations in SAE rates can be identified at all at site level using the current methods and with this size of data.

A solution to the challenges described above may be to estimate the probabilities for each site to have the number of events that they have given their individual patient times and base the KRI-analysis on these data instead of the raw AE and SAE rates. This may lead to normally distributed data without a large proportion of zero values which could possibly allow for easy setting of threshold limits. It may also allow inclusion of all current data rather than inclusion of sites with sufficient patient time. Alternatively, in terms of SAE, analysis on sites with zero reported events may be sufficient since SAE is a KRI with generally low numbers of events. However, then high levels of reported SAE would not be identified.

Taken together, a great deal of knowledge regarding a given indication in combination with the type of study is needed to set up methods for KRI-based supervised analysis in a given study. Therefore, one starting point may be to perform unsupervised analysis of previous studies on the same indication to collect sufficient data to set for example threshold limits. Naturally, another aspect would be to utilise professional experience of clinical studies and that on relevant medical indications since for example healthy volunteers or patients with mild disorders will have very different levels of AE and SAE compared to for example cancer patients.

7.2.2. Unsupervised analysis

Contrarily to supervised analysis, when performing unsupervised analysis one has the benefit of being able to use existing statistical methods. Also, the choice of methods is determined by the data at hand and no preconceptions are needed. Therefore, the set-up and performance of an unsupervised analysis for a given study may be straight forward. However, this situation proposes other challenges such as sufficient amounts of data to allow for effective and reliable statistical analysis. Overall, the current study was of a sufficiently large data set and the methods used worked well for the KRI AE. However, this was not the case for SAE where a large proportion of the data had the value zero resulting in a much skewed distribution and loss of data points. Therefore, logarithmic transformation of data was not sufficient. Still, a logarithmic transformation may be useful for AE which had a smaller proportion of zero-valued data. The large inclusion of sites for analysis on SAE might have hindered the analysis and it is plausible that probability levels of 5% for also SAE (as for AE) would be a more appropriate inclusion criteria. Again and as for supervised analysis, this problem would arise for any KRI that consists of a counting value.

The choice of statistical test is another challenge. One-way analysis of variance may not be a good choice for SAE since the residuals were only questionably normal and other tests such as non-parametric tests should be considered. The data itself was not normal and neither was that of AE. Still, scatter plots seem to be an efficient way to identify some sites with possible deviant behavior. However, a potential solution for the described challenges in terms of adverse events is the same as proposed for supervised analysis: analysis on the probabilities of rates rather than the actual rates. If this can be proposed for other KRIs as well needs to be investigated.

As in supervised analysis, inclusion criteria is a relevant challenge in unsupervised analysis. In the current study, the unsupervised analysis (both adverse events and efficacy biomarkers) was done on the same data set as for the supervised analysis and the effect of varying the inclusion was not tested. Naturally, one could argue that it makes sense to use the same inclusion criteria for adverse events regardless of analysis type. However, in terms of efficacy biomarkers this makes little sense other than out of ease if efficacy biomarkers alone are the variables to analyse. Still, in an actual RBM many KRIs

and other variables are tested and it may not be feasible to have different inclusion criteria for each sub-analysis. One reasonable proposition may be to create an over-all inclusion criterion and then perform all evaluation on this. Another may be to create parallel data sets with decreasing inclusion and find the optimal one for each KRI.

The purpose of CSM is to find signals in collected data that may indicate problems with patient safety and/or data quality and that therefore merits further investigation (TransCelerateBioPharma 2013, FDA 2017). Often a very larger number of tests are done while correction for multiple testing typically are not. This means that the risk of type I errors is great. However, this is generally not considered a problem since the outcome is not treated as evidence of findings but as indicators of possible risks of quality loss. If further investigation finds, following a significant finding of deviant data, that no risk is present then the initial finding is no longer of interest. Congruently, no corrections for multiple testing was done in this study. Still, it is reasonable to discuss which alpha levels that ought to be considered significant. CSM aims to find extremes and slight deviations are probably not important in terms of risk of quality loss. Therefore, on one hand, it ought to be reasonable that the alpha level is set at a low risk. On the other hand, there is no loss (in quality) for finding more indicators of risk than necessary. Furthermore, the outcome from an early CSM compared to a late CSM may differ at a given alpha level due to differences in data size. Hence, the alpha levels probably need to be determined depending on the circumstances for each CSM. However, another possibility would be to use different methods depending on where in a study the CSM takes place. Statistical analysis ought to become more standard towards the end of a study when large data is available while early CSM may demand specific statistical methods such as for example non-parametric tests rather than parametric.

Mahalanobis distance has previously been used with success in finding outliers and the idea is that multivariate analysis ought to reveal more outliers than univariate (Oba 2016, De Maesschalck 2000). However, the method may not have been thoroughly evaluated here since the evaluation of efficacy biomarkers did not reveal (much) deviant behaviour. It may be that the chosen biomarkers were not optimal for a method evaluation.

One problem that was experienced when computing Mahalanobis distance in multivariate analysis for several biomarkers (not reported here) was that it amplified the number of missing data. If one variable has a missing data then the multivariate Mahalanobis distance will be missing. In the current data, many biomarkers where not measured at the same visit for every patient. Without information on if these could be equivalent and possibly combined, this led to large amounts of missing data of the multivariate Mahalanobis distance.

7.3. Concluding remarks

Risk based monitoring using central statistical monitoring has previously been shown to possibly improve the monitoring process compared to source data verification (Lindblad 2014) and the implementation of it may thus increase the quality of clinical studies. Here, some of the involved aspects were used on a subsection of data from a clinical study conducted by TFS. The use of key risk indicators (KRIs) in either supervised or unsupervised analysis seems effective. The greater difficulties lie in the set-up of appropriate methods, especially for supervised analysis, and more investigations are needed for this process. KRI-based supervised analysis may seem to be a straight forward process but at least in terms of AE and SAE it proved to be rather complex. Several aspects require consideration; such as inclusion criteria, data distribution and transformations, zero-probabilities, basis for thresholds and cut-off limits for these, choice of statistical tests, alpha-levels etcetera. Unsupervised analysis using non-KRIs could not be satisfactorily evaluated here. Overall, more studies are needed that share professional experience of RBM and hands-on knowledge of the concepts to drive the enhancement of quality in clinical studies.

8. References

- Al-Marzouki S, Evans S, Marshall T, Roberts I. Are these data real? Statistical methods for the detection of data fabrication in clinical trials. *Br. Med. J.* 2005; **331**: 267-270.
- Andersen J P, Byrjalsen I, Bihlet A, Kalakou F, Hoeck H C, Hansen G, Hansen H B, Karsdal M A, Riis B J. Impact of source data verification on data quality in clinical trials: an empirical post hoc analysis of three phase 3 randomized clinical trials. *Br. J. Clin. Pharmacol.* 2014; **79** (4): 660-668
- Buyse M. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Statist. Med.* 1999; **18**: 3435-3451
- Buyse M. Centralized statistical monitoring as a way to improve the quality of clinical data. *Appl Clin Trials* 2014. Available at: http://www.appliedclinicaltrialsonline.com/centralized-statistical-monitoring-way-improve-quality-clinical-data (accessed 1 September 2017).
- De Maesschalck R, Jouan-Rimbaud D, Massart DL. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems* 2000; **50**(1): 1-18.
- eClinical Forum Risk Based Monitoring Task Force. Risk-based approached, best practices for ensuring clinical data quality. 2012 Available at (requires free registration): http://eclinicalforum.org/Downloads/tabid/59/id/87/language/en-US/Default.aspx (accessed 1 September 2017).
- FDA. Guidance to Industry Oversight of Clinical Investigations - A Risk-Based Approach to Monitoring. Available at: https://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm269919.pdf (accessed 1 September 2017).
- ICH E6 (R2) GCP Guidelines. Available at: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R2__Step_4.pdf (accessed 1 September 2017).
- Kirkwood AA, Cox T and Hackshaw A. Application of methods for central statistical monitoring in clinical trials. *Clin Trials* 2013; **10**: 783-806.
- Lindblad A S, Manukyan Z, Purohit-Sheth T, Gensler G, Okwesili P, Meeker-O´Connell A, Ball L, Marler J R. Central site monitoring: Results from a test of accuracy in identifying trials and sites failing Food and Drug Administration inspection. *Clinical Trials* 2014; **11**: 205-217.
- Oba K. Statistical challenges for central monitoring in clinical trials: a review. *Int J Clin Oncol* 2016; **21**:28–37
- Timmermans C, Venet D, Burzykowski. Data-driven risk identification in phase III clinical trials using central statistical monitoring. *Int J Clin Oncol* 2016; **21**:38-45.
- TransCelerateBioPharma Inc. Position paper: risk-based monitoring methodology. May 2013. Available at: http://www.transceleratebiopharmainc.com/wp-content/uploads/2016/01/TransCelerate-RBM-Position-Paper-FINAL-30MAY2013.pdf.pdf (accessed 1 September 2017).
- Venet D, Doffagne E, Burzykowski T, et al. A statistical approach to central monitoring of data quality in clinical trials. *Clin Trials* 2012; **9**:705-713.

9. Appendices

9.1. Background information

Appendix 1.1. Description of the clinical trial phases.

Phase	N	Subjects and doses	Primary goal
Pre-clinical	unlimited	In vitro or in vivo, not human subjects.	To test efficacy, toxicity and pharmacokinetic properties of potential drug candidates.
0	10	Healthy volunteers and (ultra)sub-therapeutic doses.	To test safety and pharmacokinetic properties; such as bioavailability and half-life, of drug candidates. Can be combined with phase I.
I	20-100	Healthy volunteers and sub-therapeutic doses (patients if cancer drugs).	To test safety and pharmacokinetic properties; such as bioavailability and half-life, of drug candidates.
II	100-300	Patients and therapeutic doses.	To determine existence of efficacy in human and to identify side effects of drug candidate.
III	300-3000	Patients and therapeutic doses.	To assess efficacy and to monitor safety of the therapeutic.
IV	unlimited	Patients that have been prescribed the pharmaceutical.	Postmarketing surveillance, to monitor safety and efficacy of the therapeutic.

9.2. Result tables and figures

Appendix 2.1. CSM – Supervised analysis; excluded sites.

Country Sites	Subjects	Patient time (days)	AE	SAE
Country -01				
All sites	4	240	0	0
Site 01	3	169	0	0
Site 02	0	.	0	0
Site 05	1	71	0	0
Country 01				
Site 02	1	213	3	0
Site 05	2	429	3	1
Site 06	3	343	0	0
Site 09	3	370	1	0
Country 02				
Site 08	1	268	0	0
Site 09	2	202	0	0
Site 10	3	296	0	0
Site 11	1	148	0	0
Country 03				
Site 03	1	174	8	0
Country 04				
Site 04	1	315	1	0
Country 05				
Site 04	2	415	0	0
Site 07	1	267	1	0
Country 06				
Site 04	0	.	0	0
Country 07				
Site 02	0	.	0	0
Site 03	4	293	3	0
Site 07	0	.	0	0
Country 08				
Site 01	3	356	1	0
Site 04	2	408	2	1
Country 09				
Site 04	3	429	0	0
Site 05	2	304	2	0
Country 10				
Site 06	1	334	2	0

AE: number of AE, SAE: number of SAE.

Appendix 2.2. CSM – Supervised analysis; adverse events summary; total, total country rates and rates by site.

	Patient time (days)	AE	SAE	Prob (no AE)	Prob (no SAE)	AE /pat.year	SAE /pat.year	SAE/AE
TOTAL	97548	685	43			2.56	0.161	6%
Country 01								
All sites	9880	19	2			0.70-Y	0.074-Y	11%
Site 01	1682	5	1			1.09-Y		20%
Site 03	6080	6	1			0.36-R	0.060-Y	17%
Site 04	775	0	0	0.48%-R				
Site 07	443	1	0			0.82-Y		0%
Site 08	900	7	0			2.84		0%
Country 02								
All sites	18851	139	11			2.69	0.213	8%
Site 01	1601	7	0			1.60		0%
Site 02	2999	15	0			1.83		0%
Site 03	2835	19	1			2.45		5%
Site 04	2004	14	1			2.55		7%
Site 05	4879	27	3			2.02	0.225	11%
Site 06	1437	9	2			2.29		22%
Site 07	3096	48	4			5.66-Y		8%
Country 03								
All sites	4077	33	3			2.96	0.269	9%
Site 01	1429	14	0			3.58		0%
Site 04	513	1	0			0.71-Y		0%
Site 06	478	1	0			0.76-Y		0%
Site 08	649	0	0	1.14%-Y				
Site 10	1008	17	3			6.16-Y		18%
Country 04								
All sites	14360	25	1			0.64-Y	0.025-R	4%
Site 01	3031	5	0			0.60-R		0%
Site 02	4963	0	0	0.00%-R				
Site 03	3272	16	1			1.79		6%
Site 05	1254	0	0	0.02%-R				
Site 06	1840	4	0			0.79-Y		0%
Country 05								
All sites	2112	22	2			3.80		9%
Site 01	812	4	1			1.80		25%
Site 02	580	7	0			4.41		0%
Site 09	720	11	1			5.58-Y		9%

Prob (no AE) and Prob(no SAE): the probability for each site or country to have no AEs or SAEs, respectively. Computations as described in the methods section. AE: number of AEs, SAE: number of SAEs, /pat.year: per patient year. Probability flags; Y (YELLOW); Prob(no AE or SAE) 1-5%, R (RED); Prob(no AE or SAE) <1%. Rate flags; Y (YELLOW): Rate value in intervals -1.0MAD to -0.5MAD or +2MAD to +4MAD from median. R (RED): Rate <1.0MAD or >4MAD from median. GREEN flags are not indicated. SAE rates were investigated only if the total patient time was sufficiently long (Prob(no SAEs) <20).

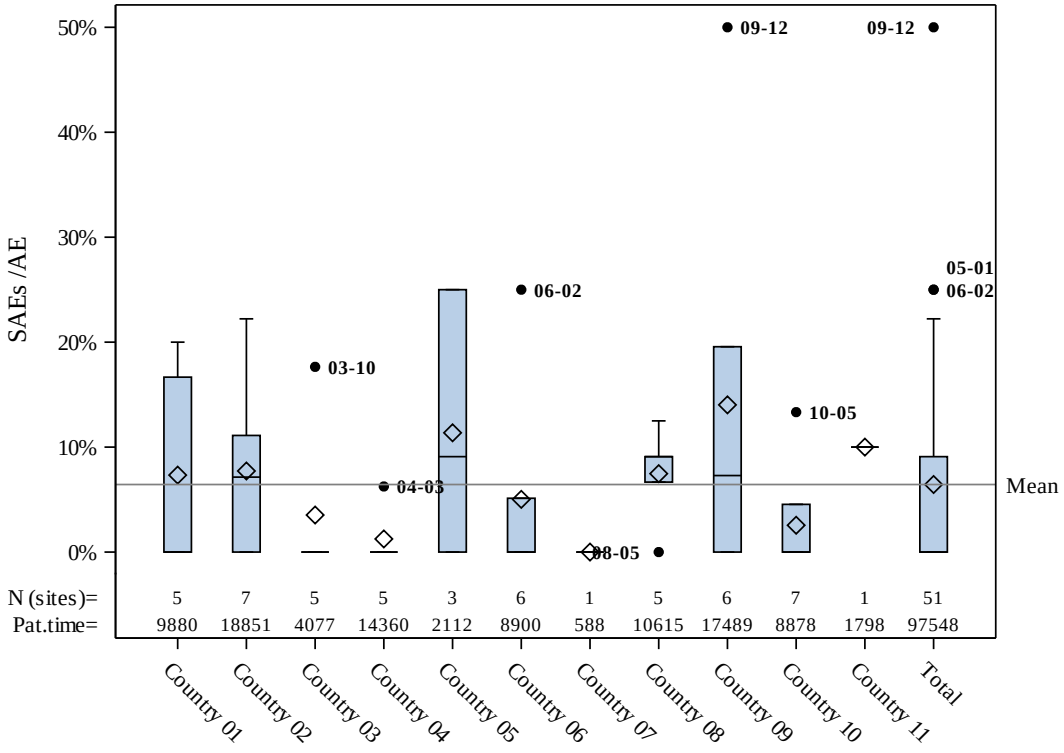
Appendix 2.2 continues on next page.

Appendix 2.2 continued.

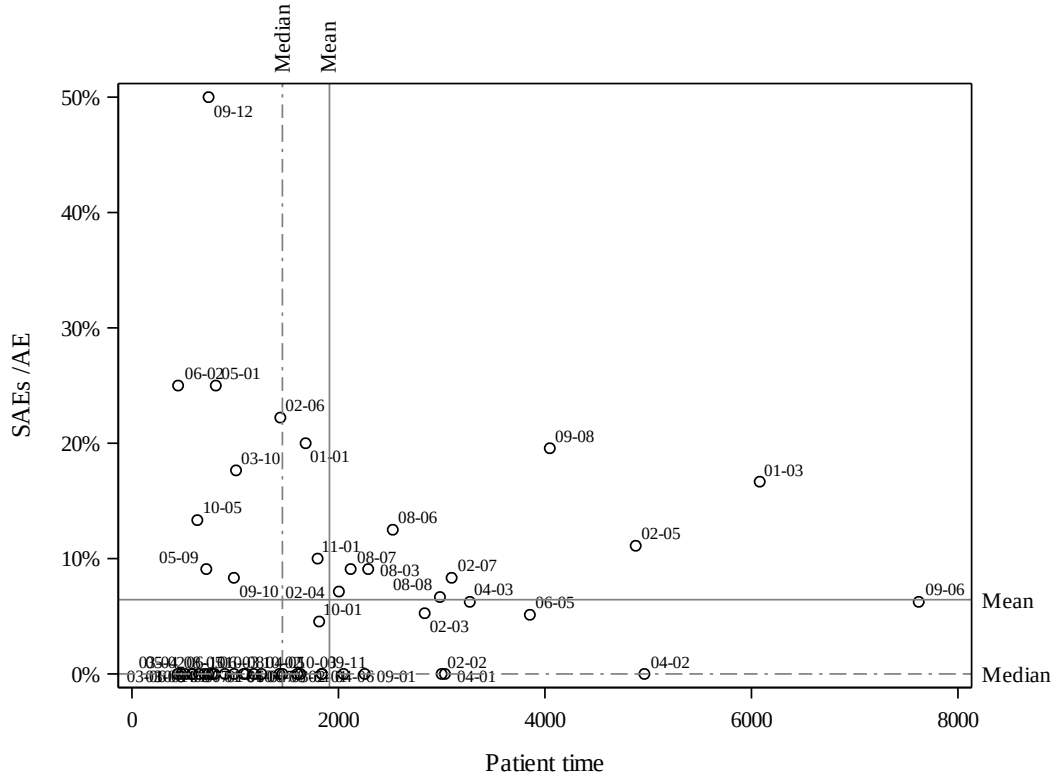
	Patient time (days)	AE	SAE	Prob (no AE)	Prob (no SAE)	AE /pat.year	SAE /pat.year	SAE/AE
Country 06								
All sites	8900	75	3			3.08	0.123	4%
Site 01	1631	9	0			2.02		0%
Site 02	447	4	1			3.27		25%
Site 03	799	6	0			2.74		0%
Site 05	3854	39	2			3.7	0.190	5%
Site 08	1181	15	0			4.64-Y		0%
Site 10	988	2	0			0.74-Y		0%
Country 07								
All sites	588	0	0	1.74%-Y				
Site 01	588	0	0	1.74%-Y				
Country 08								
All sites	10615	46	4			1.58	0.138	9%
Site 03	2288	11	1			1.76		9%
Site 05	700	1	0			0.52-R		0%
Site 06	2526	8	1			1.16-Y		13%
Site 07	2118	11	1			1.9		9%
Site 08	2983	15	1			1.84		7%
Country 09								
All sites	17489	96	12			2.00	0.251	13%
Site 01	2254	11	0			1.78		0%
Site 06	7620	16	1			0.77-Y	0.048-Y	6%
Site 08	4047	46	9			4.15	0.812-R	20%
Site 10	987	12	1			4.44		8%
Site 11	1839	9	0			1.79		0%
Site 12	742	2	1			0.98-Y		50%
Country 10								
All sites	8878	210	3			8.64-R	0.123	1%
Site 01	1813	22	1			4.43		5%
Site 02	1457	45	0			11.28-R		0%
Site 03	2049	54	0			9.63-R		0%
Site 04	739	13	0			6.43-Y		0%
Site 05	633	15	2			8.66-R		13%
Site 07	1082	41	0			13.84-R		0%
Site 08	1105	20	0			6.61-Y		0%
Country 11								
All sites	1798	20	2			4.06		10%
Site 01	1798	20	2			4.06		10%

Prob (no AE) and Prob(no SAE): the probability for each site or country to have no AEs or SAEs, respectively. Computations as described in the methods section. AE: number of AEs, SAE: number of SAEs, /pat.year: per patient year. Probability flags; Y (YELLOW); Prob(no AE or SAE) 1-5%, R (RED); Prob(no AE or SAE) <1%. Rate flags; Y (YELLOW): Rate value in intervals -1.0MAD to -0.5MAD or +2MAD to +4MAD from median. R (RED): Rate <1.0MAD or >4MAD from median. GREEN flags are not indicated. SAE rates were investigated only if the total patient time was sufficiently long (Prob(no SAEs) <20).

Appendix 2.3. CSM – unsupervised; boxplot of SAE/AE rates by country and total. Individual site values are marked (country number – site number) if outliers; i.e. values farther than 1.5*IQR (inter quartile range) from 75-percentile. Refline is mean of total.



Appendix 2.4. CSM – unsupervised; scatter plot of SAE/AE rates versus patient time (days). Reflines are mean and median of SAE/AE and Patient time, respectively.

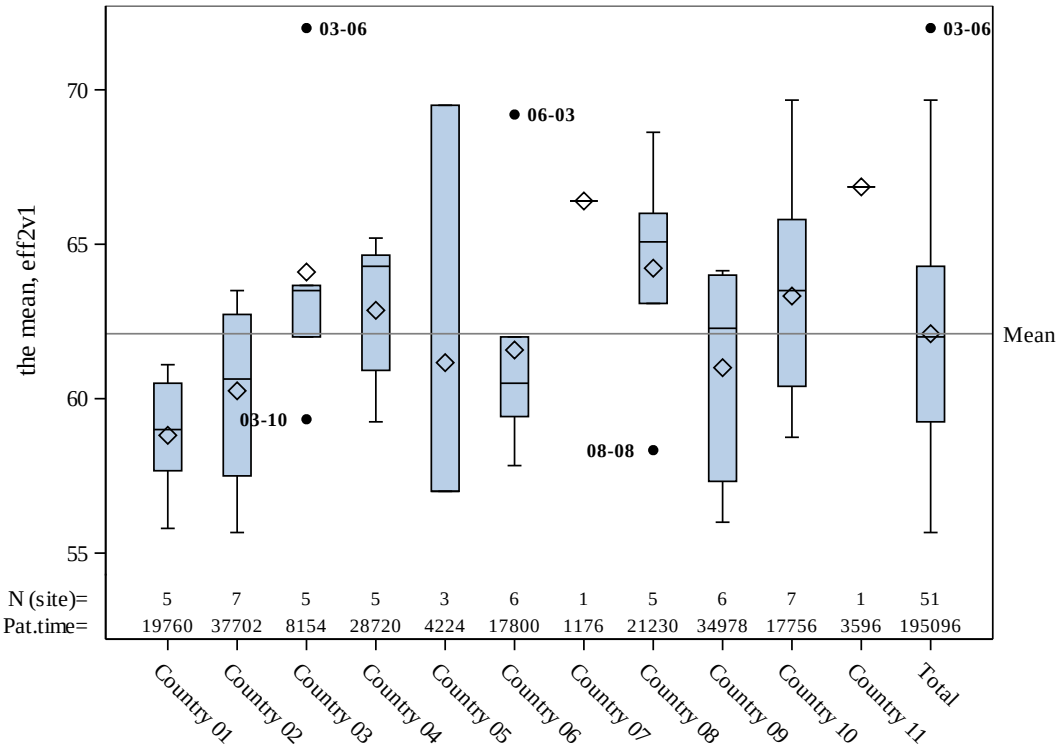


Appendix 2.5 CSM – Summary of findings from KRI-based analysis of AE and SAE. Countries and sites with some deviant data are listed. Non-bolded text is result from supervised analysis. Boxes mark findings from both supervised and unsupervised analysis. Bold text marks additional findings (outliers) from unsupervised analysis.

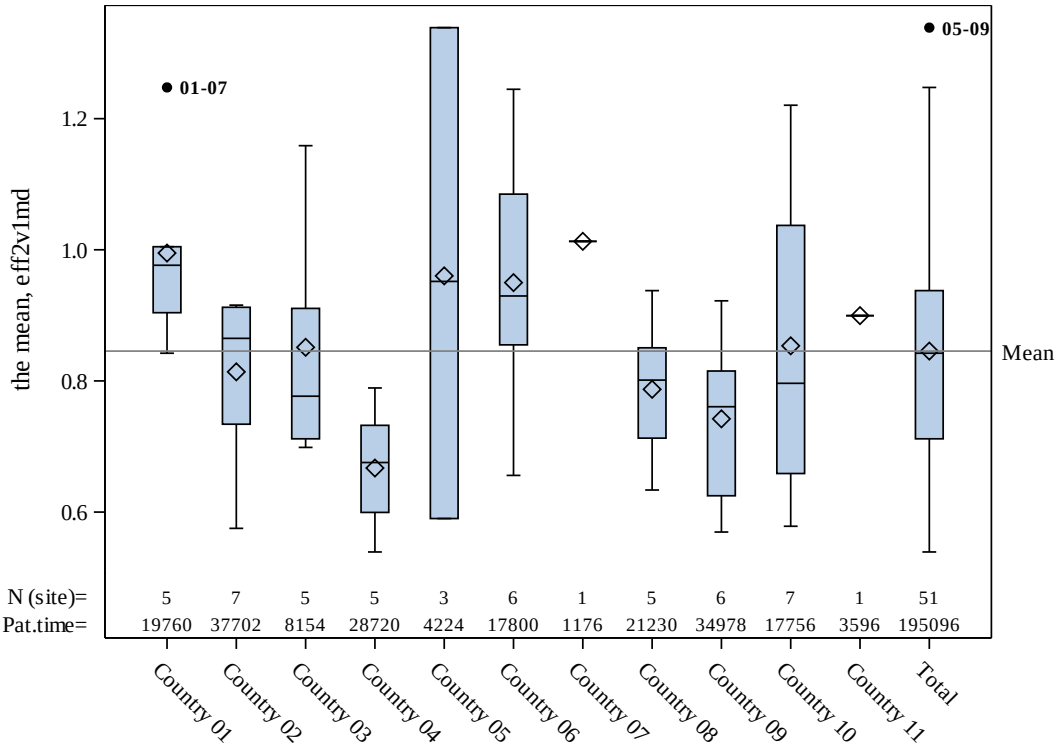
Country	Site	Patient time (days)	AE	SAE	AE /pat.year	SAE /pat.year	SAE/AE
Country 01	All sites	9880	19	2	0.70-Y	0.74-Y	
Country 01	Site 01	1682	5	1	1.09-Y	**High OUT**	
Country 01	Site 03	6080	6	1	0.36-R	0.060-Y	
Country 01	Site 04	775	0		0.48%-R		
Country 01	Site 07	443	1	0	0.82-Y		
Country 01	Site 08	900	7	0	**High OUT**		
Country 02	Site 02	2999	15	0		**Low OUT**	
Country 02	Site 07	3096	48	4	5.66-Y		
Country 03	Site 04	513	1	0	0.71-Y		
Country 03	Site 06	478	1	0	0.76-Y		
Country 03	Site 08	649	0		1.14%-Y		
Country 03	Site 10	1008	17	3	6.16-Y	**High OUT**	**High OUT**
Country 04	All sites	14360	25	1	0.64-Y	0.025-R	
Country 04	Site 01	3031	5	0	0.60-R		
Country 04	Site 02	4963	0		0.00%-R		
Country 04	Site 03	3272	16	1		**High OUT**	**High OUT**
Country 04	Site 05	1254	0		0.00%-R		
Country 04	Site 06	1840	4	0	0.79-Y		
Country 05	Site 01	812	4	1			**High OUT**
Country 05	Site 09	720	11	1	5.58-Y		
Country 06	Site 02	447	4	1		**High OUT**	**High OUT**
Country 06	Site 08	1181	15	0	4.64-Y		
Country 06	Site 10	988	2	0	0.74-Y		
Country 07	Site 01	588	0		1.74%-Y		
Country 08	Site 05	700	1	0	0.52-R	**Low OUT**	**High OUT**
Country 08	Site 06	2526	8	1	1.16-Y		
Country 09	Site 06	7620	16	1	0.77-Y	0.048-Y	
Country 09	Site 08	4047	46	9		**High OUT**	
Country 09	Site 12	742	2	1	0.98-Y		**High OUT**
Country 10	All sites	8878	210	3	8.64-Y		
Country 10	Site 02	1457	45	0	11.28-R		
Country 10	Site 03	2049	54	0	9.63-R		
Country 10	Site 04	739	13	0	6.43-Y		
Country 10	Site 05	633	15	2	8.66-R	**High OUT**	**High OUT**
Country 10	Site 07	1082	41	0	13.84-R		
Country 10	Site 08	1105	20	0	6.61-Y		

AE: number of AE, SAE: number of SAE, /pat.year: per patient year. Y; yellow flag, R; red flag. OUT: outlier. For sites with zero AE, the probability to have zero AE is given instead of AE rate.

Appendix 2.6. CSM – Unsupervised analysis; Efficacy biomarker 2 (eff1v1) across countries; means for each site. Number (N) of patients and sites are given for each country, respectively. Individual site values are marked (country number – site number) if outliers; i.e. values farther than 1.5*IQR (inter quartile range) from 75-percentile. Refline is mean of total.



Appendix 2.7 CSM – Unsupervised analysis; Mahalanobis distance for efficacy biomarker 2 (eff2v1md) across countries; means for each site. Number (N) of patients and sites are given for each country, respectively. Individual site values are marked (country number – site number) if outliers; i.e. values farther than 1.5*IQR (inter quartile range) from 75-percentile. Refline is mean of total.



9.3. SAS® Code

Appendix 3.1. Code to estimate mean of the assumed poisson distributions, for AE and SAE, respectively; to exclude sites with to small patient time.

```
%MACRO sum(VAR=);
proc univariate data=sitetemp3 noprint;
                var &var;
                output out=sum sum=&var.S;
run;
%global &var.S;
proc sql noprint;
                select &var.S into :&var.S from sum;
quit;
%MEND;
%sum(VAR=Nae);
%sum(VAR=studtm);
%sum (VAR=Nsae);
data lambdadata;
                lambdaAE=(&naeS/&studtmS);
                lambdaSAE=(&nsaeS/&studtmS);
run;
data lambda;
                set lambdadata;
                t_5p=-log(0.05)/lambdaAE;
                t2_20p=-log(0.2)/lambdaSAE;
run;
%MACRO globalmacrovar(VAR=);
%global &var;
data _null_;
                set lambda;
                %let &var="&var";
                call symput("&var",&var);
run;
%MEND;
%globalmacrovar(VAR=t_5p);
%globalmacrovar(VAR=t2_20p);
data ad.excluded;
                set ad.site;
                if studtm<&t_5p;
run;
```

Appendix 3.2. Code to compute and flag threshold analysis.

```
%MACRO threashold(VAR=,LIMIT=);
proc univariate data=aeinclude noprint;
                    where studtm>&limit and site ne 'TOTAL' and site ne 'All sites';
                    var &var;
                    output out=stat median=median mad=mad;
run;
data _null_;
                    set stat;
                    %let x="&var.median";
                    call symput("x",median);
                    %let mad="&var.mad";
                    call symput("mad",mad);
run;
data &var.result;
                    set aeinclude;
                    format &var.THR $char6.;
                    if &var=0 then &var.THR='.';
                    else if &var=" then &var.THR='.';
                    else if studtm<&limit then &var.THR='N/A';
                    else if &var<=&x+2*&mad and &var>=&x-0.5*&mad
                                                then &var.THR='GREEN';
                    else if &var>&x+2*&mad and &var<=&x+4*&mad
                                        or &var>=&x-1*&mad and &var<&x-0.5*&mad
                                                then &var.THR='YELLOW';
                    else if &var>&x+4*&mad or &var<&x-1*&mad
                                                then &var.THR='RED';
                    else &var.THR='.';
                    keep site country id &var.THR;
run;
proc sort data=ad.aeincluded; by id; run;
proc sort data=&var.result; by id; run;
data ad.aeincluded;
                    merge ad.aeincluded(in=start) &var.result(in=result);
                    by id;
run;
%MEND;
%threashold(VAR=naer,LIMIT=&t_5p);
%threashold(VAR=nsaer,LIMIT=&t2_20p);
```

Appendix 3.3. Code to compute Mahalanobis distance, univariate and multivariate, respectively.

Univariate:

```
%MACRO mahalanobis(IN_OUTFILE=,VAR=);
proc princomp data=&in_outfile std out=out outstat=outstat noprint;
   var &var;
run;
data mahala;
   set out;
                  &var.md = sqrt(uss(of prin:));
                  drop prin1;
run;
proc sort data=&in_outfile; by subjid; run;
proc sort data=mahala; by subjid; run;
data &in_outfile;
                  merge &in_outfile mahala;
                  by subjid;
run;
%MEND;
%mahalanobis(IN_OUTFILE=ad.effwide,VAR=eff1v1);
%mahalanobis(IN_OUTFILE=ad.effwide,VAR=eff2v1);
```

Multivariate:

```
%MACRO multi_mahalanobis(IN_OUTFILE=,VAR1=,VAR2=);
proc princomp data=&in_outfile std out=out outstat=outstat noprint;
   var &var1 &var2;
run;
data mahala;
   set out;
                  &var1.&var2.md = sqrt(uss(of prin:));
                  drop prin:;
run;
proc sort data=&in_outfile; by subjid; run;
proc sort data=mahala; by subjid; run;
data &in_outfile;
                  merge &in_outfile mahala;
                  by subjid;
run;
%MEND;
%multi_mahalanobis(IN_OUTFILE=ad.effwide,VAR1=eff1v1,VAR2=eff2v1);
```