Drug Saf (2014) 37 (Suppl 1):S33–S45
DOI 10.1007/s40264-014-0184-5

REVIEW ARTICLE

Methodology to Assess Clinical Liver Safety Data

Michael Merz · Kwan R. Lee · Gerd A. Kullak-Ublick · Andreas Brueckner · Paul B. Watkins

Abstract Analysis of liver safety data has to be multivariate by nature and needs to take into account time dependency of observations. Current standard tools for liver safety assessment such as summary tables, individual data listings, and narratives address these requirements to a limited extent only. Using graphics in the context of a systematic workflow including predefined graph templates is a valuable addition to standard instruments, helping to ensure completeness of evaluation, and supporting both hypothesis generation and testing. Employing graphical workflows interactively allows analysis in a team-based setting and facilitates identificatio\n of the most suitable graphics for publishing and regulatory reporting. Another important tool is statistical outlier detection, accounting for the fact that for assessment of Drug-Induced Liver Injury, identification and thorough evaluation of extreme values has much more relevance than measures of central tendency in the data. Taken together, systematical graphical data exploration and statistical outlier detection may have the potential to significantly improve assessment and interpretation of clinical liver safety data. A workshop was convened to discuss best practices for the assessment of drug-induced liver injury (DILI) in clinical trials.

Key Points

In addition to standard summary tables and narratives, graphics can help significantly to improve liver safety assessment

A systematic workflow helps to ensure completeness of evaluations and supports hypothesis generation and testing

To differentiate true outliers from random variation, robust statistical methods are available that should be considered for liver safety evaluation

M. Merz (✉) · G. A. Kullak-Ublick
Discovery and Investigative Safety, Novartis Institutes for BioMedical Research, Klybeckstrasse 141, WKL-135.1.78, 4057 Basel, Switzerland
e-mail: michael.merz@novartis.com

K. R. Lee
Medical Analytics, GlaxoSmithKline, Collegeville, PA, USA

G. A. Kullak-Ublick
Department of Clinical Pharmacology and Toxicology, University Hospital Zurich, Zurich, Switzerland

A. Brueckner
Novartis Pharma AG, Basel, Switzerland

P. B. Watkins
The Hamner-University of North Carolina Institute for Drug Safety Sciences, Research Triangle Park, NC, USA

P. B. Watkins
Schools of Medicine, Pharmacy and Public Health, University of North Carolina, Chapel Hill, NC, USA

1 Introduction

Timely detection and proper assessment of drug-induced liver injury (DILI) in clinical trials has for decades been one of the key safety challenges for both pharmaceutical industry and regulatory authorities.

A workshop was sponsored and organized jointly by the European Innovative Medicines Initiative (IMI) and the

Hamner Institute for Drug Safety Sciences (IDSS), with the aim of addressing gaps in current guidance and initiating alignment of liver safety assessment on a global scale.

On November 9, 2012, in Boston, regulatory experts from the FDA, EMA, Health Canada, and the Japanese National Institute of Health Sciences discussed with representatives from industry and academia what could be considered best practices in clinical liver safety assessment. The best practices workshop focused on four key areas: 1) data elements and data standards, 2) methodologies to systematically analyze liver safety data, 3) tools and methods for causality assessment, and 4) liver safety assessment in special populations such as hepatitis and oncology patients.

This section summarizes current methods for systematic assessment of liver safety data, as discussed at the workshop, and provides respective recommendations for use in clinical drug development.

Assessment of liver safety data needs to take into account not only classic safety biomarkers such as standard liver tests alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), and total bilirubin (TBIL), but also patient demographics, medical history, adverse events and concomitant medication. Moreover, time dependence of and covariation between liver test results have to be factored in. Thus, proper evaluation of liver safety profiles can be a highly complex task, requiring comprehensive datasets and suitable analysis methods. Standard approaches such as use of tabular summaries, narratives, and descriptive statistics may be supplemented by graphical displays in a systematic workflow and outlier detection methods.

2 Tabular Summaries

2.1 Incidence Tables

Tabular summaries may be most useful to capture and compare *incidences* of liver test elevations as well as *extent of changes from baseline* across treatment groups both at study and program level. In terms of incidences, common thresholds to capture and assess liver test related events are $>3 \times$ ULN, $>5 \times$ ULN, $>10 \times$ ULN for aminotransferase activities, and $>2 \times$ ULN for bilirubin concentrations [1].

As for aminotransferase activities, both ALT and AST are usually captured, although the added value of listing AST in addition to ALT may be limited to helping with differential diagnosis, i.e. differentiating muscle-related from liver-related ALT elevations, as well as alcohol from non-alcohol-associated etiology of liver test elevations [2, 3]. Addition of GGT, which seems to be more common in Europe as compared to the US, may increase sensitivity for cholestatic liver injury at the cost of decreasing specificity. A general recommendation to either exclude or include GGT measurements into the panel for liver safety assessment cannot be given at this point in time [4].

2.2 Shift Tables

Shift tables listing number/percentage of patients shifting e.g. from normal to above or below ULN while on treatment as compared to baseline are widely used for safety assessment in drug development. They provide a quick overview on gross changes that might be treatment related. However, a lot of valuable information may be lost by data reduction using shift tables only. A more efficient approach may be use of scatter plots displaying shifts from baseline by study visit or maximum shifts from baseline during the study, as outlined in Sect. 3.2.4.

2.3 Descriptive Statistics

Analyzing liver safety data often makes use of comparing mean and/or median changes from baseline for liver test results across treatments along with measures of variability (standard deviation, standard error, and range). Although this approach can add to understanding of drug effects on the liver, it disregards the fact that predominant interest when assessing liver safety data will be on outliers in the data. It is rather rare cases of idiosyncratic DILI than more frequent dose-dependent intrinsic DILI cases that give reason for concern and need thorough data work-up, paying attention to all individual data as well as trends in the overall dataset, association with concomitant medication, medical history, and adverse events.

Using graphics as an add-on to tabular summaries can help to address these requirements and compensate for the short-comings of the latter [5].

3 Graphical Workflows

Graphics, ideally in the setting of a defined, systematic workflow using interactive graphics software, can take into account the entirety of individual patient data as well as trends across the population, and help paying attention to the multivariate nature of safety signals and time dependency of observations. A graphical workflow can help to maximize knowledge gain from the data available, and at the same time ensure completeness of safety evaluation. Of great importance though is adhering to best practices for graphical data exploration as outlined e.g. by Tukey and Cleveland [6–8].

3.1 Prerequisites

3.1.1 Normalization

Adequately assessing liver safety data needs comparison between different continuous variables, across different studies, different laboratories, etc. To facilitate that, normalization is helpful. Simple normalization of dividing raw liver test values by the Upper Limit of Normal (ULN) values is used most often, although there are some limitations associated with this approach. Even after normalization, ULN corrected data may not be perfectly comparable across different laboratories and the associated variability might actually be misleading, possibly due to the fact that the "standard" for calculating ULN is not consistent. This has been illustrated using extensive Phase II–IV clinical trial data from a generally healthy patient population [9].

Normalization by individual baseline values may be a better alternative when data have been generated across different labs since it can reduce unnecessary variation [10–12] and is more consistent across labs. However, given that as yet there are only limited data available across different populations on use of change from baseline as compared to use of multiples of ULN, application of both normalization approaches in parallel should be considered. Graphs presented in this section are based on the as yet still more common approach of multiples of ULN, but can mostly be applied to baseline-corrected data as well. Using multiples of baseline, scatter plots of shift from baseline as presented in Sect. 3.2.4, however, could be replaced by simple box plots across study visits and parameters.

In this context, attention needs to be given to definition of baseline. As indicated in other sections of this paper, taking just one measurement as an individual patient's baseline is not adequate, given within-subject variation of liver tests [10, 12]. A more suitable determination of baseline may consist of two measurements at least two weeks but not more than two months apart. Data analysis when at least two baseline measurements are available may use minimum baseline to minimum post baseline and maximum baseline to maximum post baseline changes to account for within subject variation [13].

3.1.2 Data Types

Analyzing raw or normalized biomarker values only may be insufficient to see the complete picture. Derived variables such as absolute and relative changes from baseline, maximum values on treatment, flags for exceeding predefined threshold values etc. may be required to adequately interpret liver test results. Thus, before starting liver safety data exploration, a set of derived variables should be defined and calculated.

3.1.3 Data Structure

Typically, datasets for safety analysis include study identifier, subject identifier, visit numbers and visit names, parameter names, parameter results, lower and upper limits of normal ranges, units, and relevant covariates such as age, gender, BMI, and ethnicity, displayed by column. For most of the graphics used for liver safety exploration, this structure is sufficient. However, in order to address specific questions such as shape of bivariate distributions, shifts from baseline by visit etc., transposing the dataset by parameter names or by visits, i.e. having parameter names or visit numbers as column headers may be necessary. In order to support an efficient workflow, it is helpful to define individual steps of the workflow and required data structures upfront and make sure analysis datasets are available in all formats required.

3.1.4 Key Questions to be Addressed

Key questions to address when analyzing liver safety data comprise:

- Are there any true Hy's law cases in the dataset?
- How are changes across different liver tests correlated, and how do those correlations differ between treatment groups?
- What is the time dependent incidence of elevations of liver tests in active treatment and comparator arms? Is there a "window of susceptibility" in the active treatment arm?
- Are shifts from baseline different between treatment groups?
- Is there any evidence for a dose-response-relationship?
- What do time profiles of individual liver tests or liver test panels look like?
- Are liver test changes observed during treatment transient or progressing while a patient is on treatment?
- What do time profiles look like after treatment is stopped?
- How does intake of certain concomitant medications or occurrence and/or resolution of certain adverse events relate to time profiles of liver tests?
- Are liver test elevations correlated with the desired therapeutic effect of the drug?
- Are liver test elevations associated with non liver side effects or laboratory abnormalities?
- Are liver test elevations associated with pharmacokinetic parameters of the drug (if available)?

To systematically address these questions, a set of standard graph templates can be used and customized as required.

3.2 Graph Templates and Systematic Workflow

3.2.1 Correlations

Assessing correlations between liver tests, both absolute values on treatment and changes from baseline, can provide insight into underlying pathology of treatment associated liver effects. Exploring relationships of liver test changes with key covariates such as age, body mass index, gender, ethnicity, may help to identify risk factors for DILI.

3.2.1.1 eDISH The key graphical representation to assess a drug's liver safety profile and to immediately identify cases of special concern is the "eDISH" (evaluation of Drug-Induced Serious Hepatotoxicity plot [14]), a log/log display of correlation between peak TBIL vs. ALT, both in multiples of ULN, with horizontal and vertical lines indicating Hy's law thresholds, i.e. ALT = 3 × ULN and total bilirubin = 2 × ULN. The eDISH plot makes immediately evident subjects potentially matching Hy's law laboratory criteria, all located in the upper right quadrant of the graph. Data points in the lower right quadrant, i.e. exceeding 3 × ULN for ALT, but being below 2 × ULN for total bilirubin, suggest an increased risk for liver injury as well, if incidence is differing between active treatment and control groups, however, not to the same extent and with less specificity as compared to Hy's law.

Figure 1 shows an example of an eDISH plot, comparing pooled study drug against control data. Horizontal and vertical lines indicate Hy's law thresholds.

Patients with active treatment show a higher incidence of values in the lower right quadrant and twelve *potential* Hy's law cases in the upper right quadrant, thus suggesting a potential risk for severe drug-induced liver injury associated with this drug. As stated in the FDA's guidance on Drug-Induced Liver Injury, "…Finding one Hy's Law case in the clinical trial database is worrisome; finding two is considered highly predictive that the drug has the potential to cause severe DILI when given to a larger population." [1].

A limitation with the standard eDISH plot is its lack of displaying sequence of maximum observed values for ALT and bilirubin, i.e. which of both was first, as well as length of time intervals between maximum observed values. However, from a clinical perspective, these data are highly relevant, since only bilirubin elevations *simultaneous with or* soon *following* peak ALT elevations may indicate loss of hepatic function due to liver injury. Moreover, a long time interval, exceeding four weeks, between both peaks may also speak against a causal correlation.

Another limitation of the standard eDISH plot is its lack of displaying levels of ALP at the time of peak ALT elevation. Elevation of ALP >2 × ULN or a ratio R ([ALT × ULN]/[ALP × ULN]) <5 preceding or simultaneous with

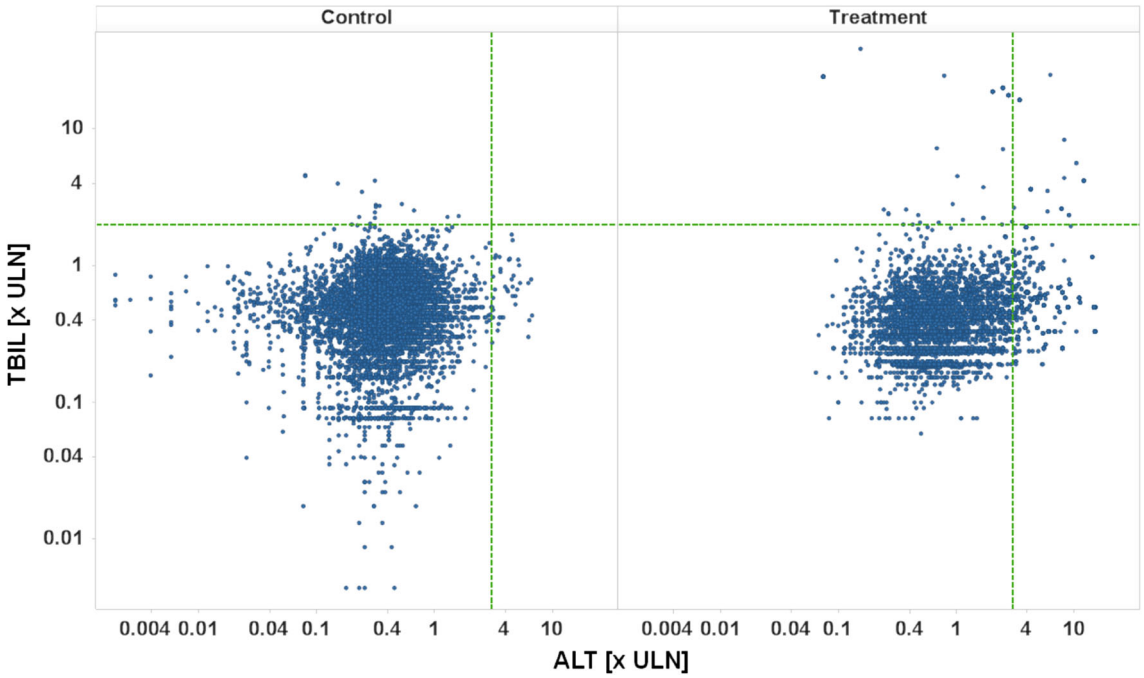

Fig. 1 eDISH plot, TBIL [× ULN] vs. ALT [× ULN] on a log/log scale, treatment by panel, pooled active versus control. *ULN* upper limit of normal, *ALT* alanine aminotransferase, *TBIL* total bilirubin

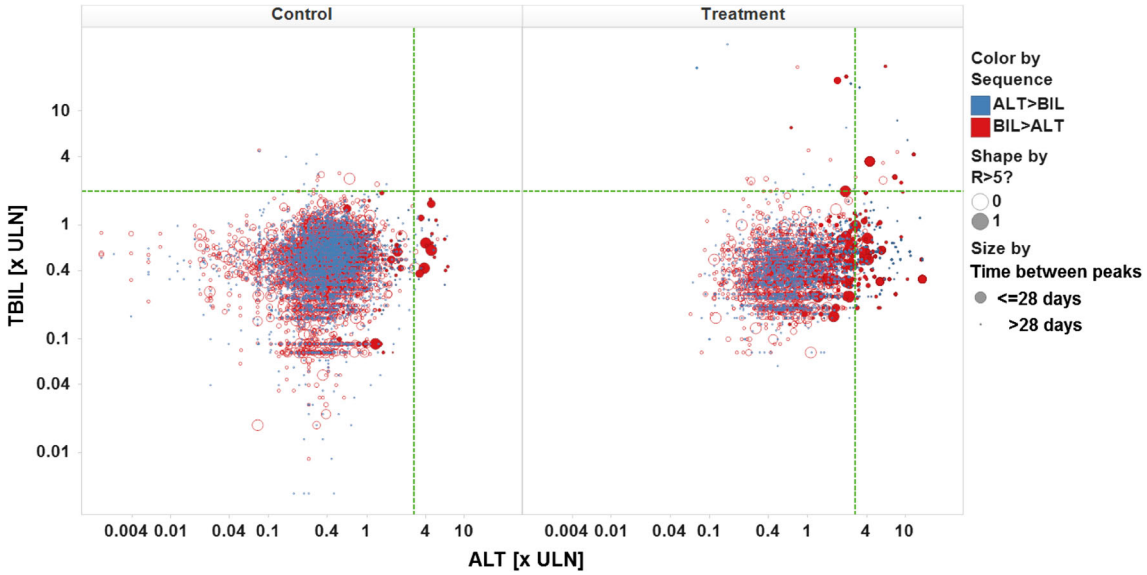

Fig. 2 Modified eDISH plot, color by sequence of peak values, size by 1/time interval between peaks, shape by R flag. *ULN* upper limit of normal, *ALT* alanine aminotransferase, *TBIL* total bilirubin, *BIL* bilirubin

ALT elevations suggests cholestatic/mixed type liver disease, such that cases of combined bilirubin and ALT elevations would not qualify as Hy's law cases.

It is desirable to have all this information included in the graphical display, as well. Figure 2 shows the same data as above, using a proposed modification to the eDISH plot, with color coding for sequences of ALT and bilirubin peaks, and size coding for the time interval between both peaks. In order to make the most relevant data points easily visible, the more concerning sequence of bilirubin parallel to or following ALT peak is coded in red, the time interval is coded as 1/interval to make shorter time intervals being displayed as larger markers. Filled circles refer to data having R >5. Thus, the data points to watch out for, qualifying as potential Hy's law cases, primarily are large, filled, red circles in the upper right quadrant.

In the above example, nine of the twelve data points in the Hy's law quadrant for patients in active treatment groups show the sequence of interest, i.e. bilirubin following or simultaneously elevated with ALT peak, but only one of those has a time interval of less than four weeks between both peaks and R >5. Thus, using this modified version of the eDISH plot, eleven out of twelve potential Hy's law cases can be immediately identified as being likely less relevant.

However, given that the eDISH plot is using peak values only, even for patients displaying a likely less relevant data point in the modified eDISH plot there may be other measurements during a trial not being peak values, but meeting Hy's law criteria, having the proper sequence of events, i.e. TBIL following ALT within a short period of time, plus an R >5. Those data could be "masked" by less relevant peak values and hence not be displayed in the eDISH plot. Thus, it needs to be underlined that, using a modified eDISH plot as outlined above, with color, shape and size coding to identify the likely more relevant cases, can only *aid prioritization* of cases but not replace thorough evaluation of *all patients displayed in the Hy's law quadrant*.

Another useful modification of eDISH takes into account changes from baseline instead of absolute values for TBIL and ALT, along with population specific thresholds, as suggested by Lin et al. [15] which is described in more detail in Sect. 4.1.1.

3.2.1.2 Other Correlations Other correlations of interest when exploring liver safety profiles of new drugs are those between different liver enzymes, i.e. ALT/AST, ALT/GGT, and ALT/ALP. Whereas in the healthy liver, ALT and AST are closely correlated, ALT and the two other enzymes usually are not. However, in some cases of DILI, elevations of ALP and/or GGT may correlate with increased ALT activities, providing some hints about the underlying pathology, i.e. cholestasis or mixed type cholestatic/hepatocellular injury. Isolated elevations of GGT activities without associated ALT or ALP changes may sometimes indicate enzyme induction rather than cell injury, as observed e.g. in cases of chronic alcohol abuse [16].

3.2.2 Time Profiles

Changes of liver tests over time can provide crucial information on both underlying pathology and causal relationship to drug treatment. Line plots of either

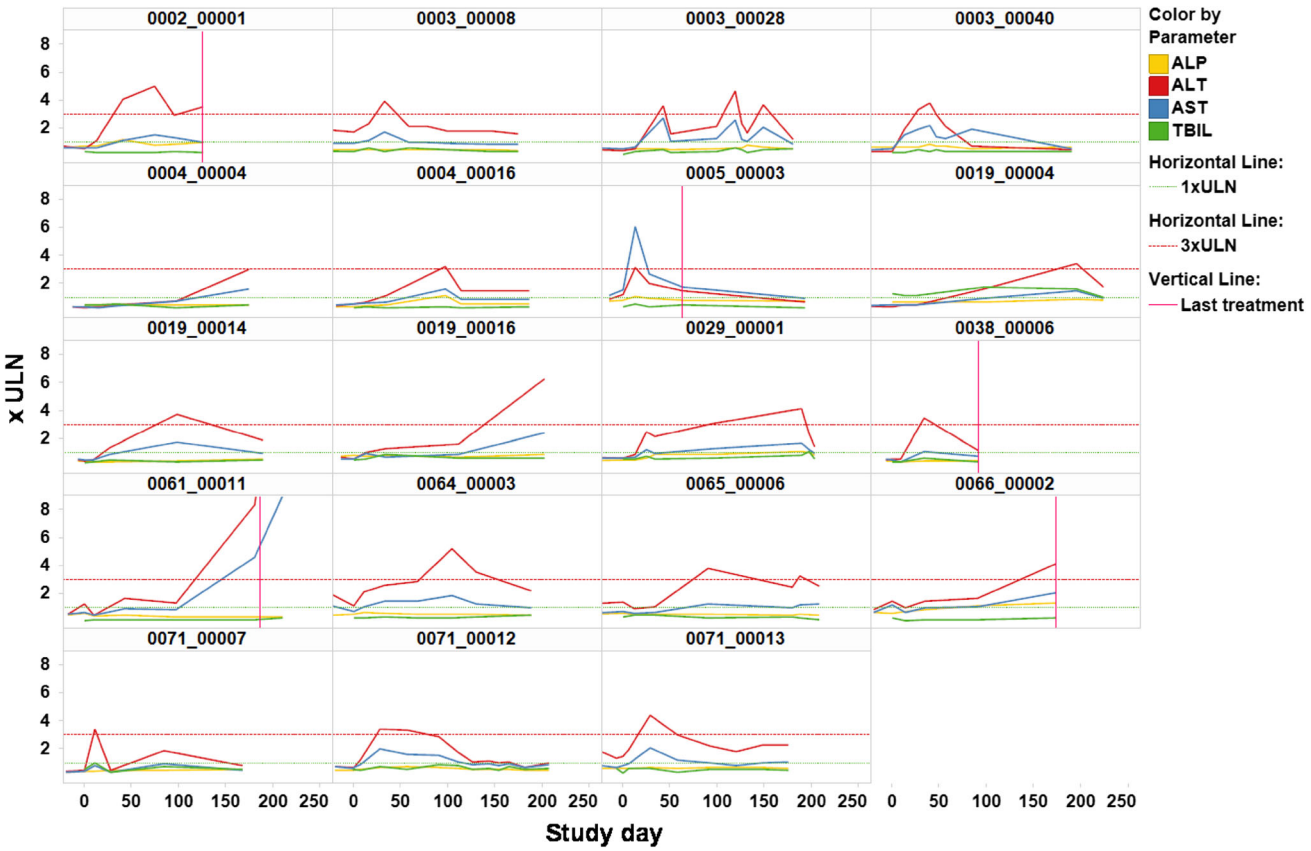

Fig. 3 Time profiles of ALT, AST, ALP, and TBIL, panel by patients, treatment end indicated by *vertical red line*, ULN and 3 × ULN indicated by *horizontal green* and *red line*, respectively. Color coding by liver test. *ALT* alanine aminotransferase, *AST* aspartate aminotransferase, *ALP* alkaline phosphatase, *TBIL* total bilirubin, *ULN* upper limit of normal

individual markers or marker panels are most useful to assess biomarker time profiles, particularly if combined with elements indicating start and/or end of drug treatment, dose levels etc. In the context of a systematic workflow using interactive graphics software, evaluation of time profiles ideally follows assessment of the eDISH plot via drill-down from selected data points, e.g. points in the Hy's law quadrant.

Figure 3 provides an overview on liver test profiles over time for 19 patients in a clinical study who showed ALT elevations >3 × ULN while on treatment, one panel per patient. Treatment end is indicated by a vertical red line, color coding is by liver test, horizontal lines represent ULN (green line) and 3 × ULN (red line), respectively.

As displayed in the plot, several patients showed short-lived, transient peaks of ALT, with serum activities decreasing despite continued treatment. Only few patients had to be taken off treatment due to continuous or worsening elevations of ALT.

Moreover, the plot allows assessing time-wise association of different biomarker effects by patient. Only patient 0004_00016 shows discrete elevation of ALP in parallel with peak ALT and AST, pointing towards a possible cholestatic component of liver injury. There are no apparent elevations of bilirubin parallel or subsequent to ALT elevations in any of the patients, confirming the rather benign nature of liver enzyme changes observed in the study.

3.2.3 *Association with Concomitant Medication and Adverse Events*

A particularly helpful graph to analyze association of liver test changes with adverse events and concomitant medication is a patient profile, defined as synoptic presentation of line plots for all three items along a shared time axis.

Figure 4 displays for an individual patient the ALT profile over time on top of the plot, concomitant medication and adverse events beneath. The horizontal red line in the top plot represents 3 × ULN for ALT. In the two lower plots, start and end times of concomitant medication intake and adverse events are displayed as blue triangles, the black line between associated triangles indicates ongoing concomitant medication or adverse event, respectively.

As displayed in the plot, the patient had taken an acetaminophen-containing medication, Nyquil™, and an

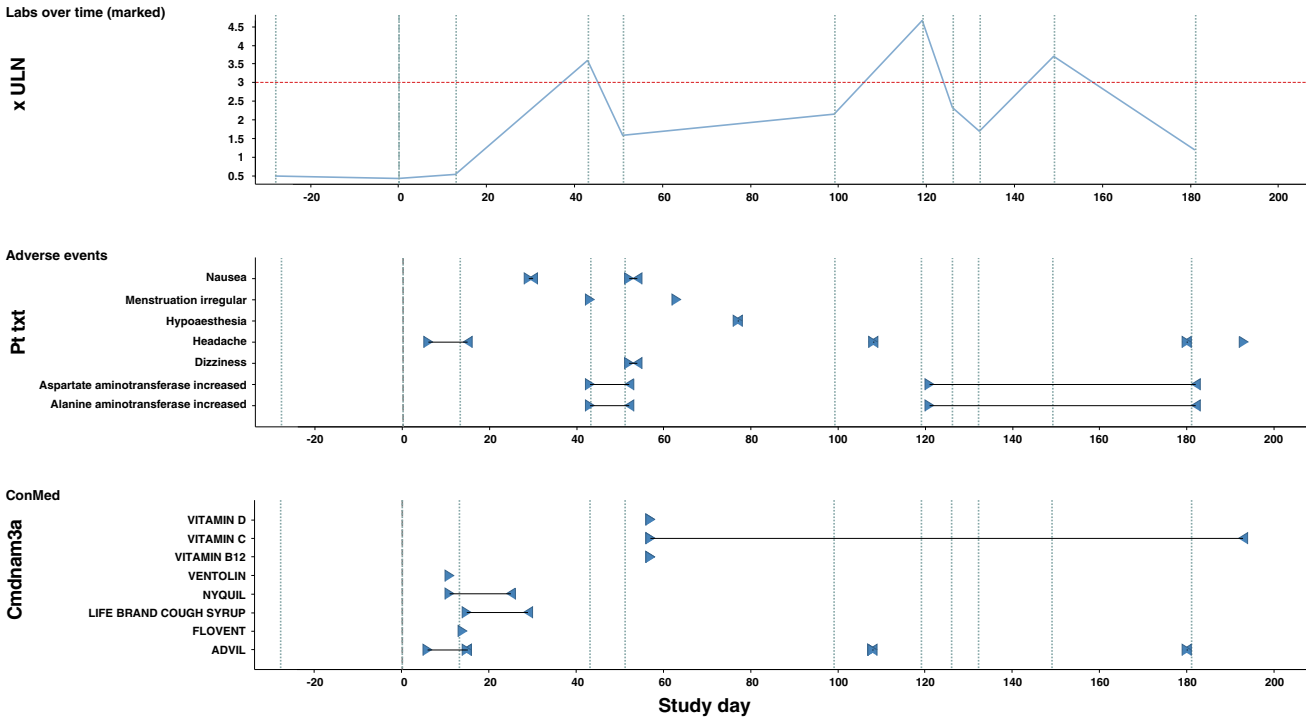

Fig. 4 Patient profile of ALT, adverse events, and concomitant medication; start and end date of adverse events and concomitant medications indicated by *blue triangles*. *ALT* alanine aminotransferase, *ULN* upper limit of normal, *ConMed* concomitant medication

ibuprofen-containing medication, Advil™, before the first ALT peak, and again Advil™, around the time of the second ALT peak. Both drugs might have been causally related to the ALT elevation. Moreover, the patient reported several adverse events of headache during the trial, one particularly preceding the second ALT peak. It is conceivable that the patient might have taken e.g. acetaminophen to treat his headache but forgotten to report this as concomitant medication.

Observing this kind of temporal association of ALT peaks with headache or other pain events in a graphical display can trigger focused re-questioning the patient to ensure no suspicious comedication has been used around the time of liver enzyme elevation.

Thus, a synoptic display of ALT profiles, concomitant medication and adverse events may sometimes help substantially to identify causes for clinically relevant changes in liver safety biomarkers.

3.2.4 Shifts from Baseline

Liver test results always have to be viewed in the context of their respective baselines to allow adequate assessment of treatment or disease effects. This can be done either by analyzing absolute and relative changes from baseline, or by using scatter plots with baselines on the x-axis and e.g. maximum post-baseline values on the y axis. When plotting only *maximum* post-baseline values on the y-axis, however, careful consideration needs to be given to the number of post-baseline measurements per patient, particularly when no control groups are available for comparison across treatments: the larger the number of post-baseline observations per patient, the more biased the plot will be towards values increasing from baseline. To avoid that, an alternative is e.g. to plot *all* post-baseline values per patient, instead of selecting the maximum values only, or displaying shifts as scatter plots by visit.

Figure 5 shows a respective example with four post-baseline observations per patient. This is a Trellis plot with treatment groups across rows and biomarker names across columns. Color coding is by gender. The blue diagonal line in each panel represents the line of identity, i.e. each value on the line corresponds to maximum post-baseline equaling baseline, each point above the line is an increase, points below the line are a decrease from baseline, respectively.

In addition, the plot allows to assess the number of patients exceeding certain threshold values, represented by the green (=ULN) and red (3 × ULN) horizontal and vertical broken lines in each panel.

In this example, there is a clear trend for higher shifts from baseline, i.e. elevations, for ALT and AST in both active treatment groups. However, even the placebo group displays some elevations from baseline at least for ALT. Although this is a phenomenon not uncommon in clinical studies and may be explained by effects of diet, physical exercise, concomitant disease or comedication,

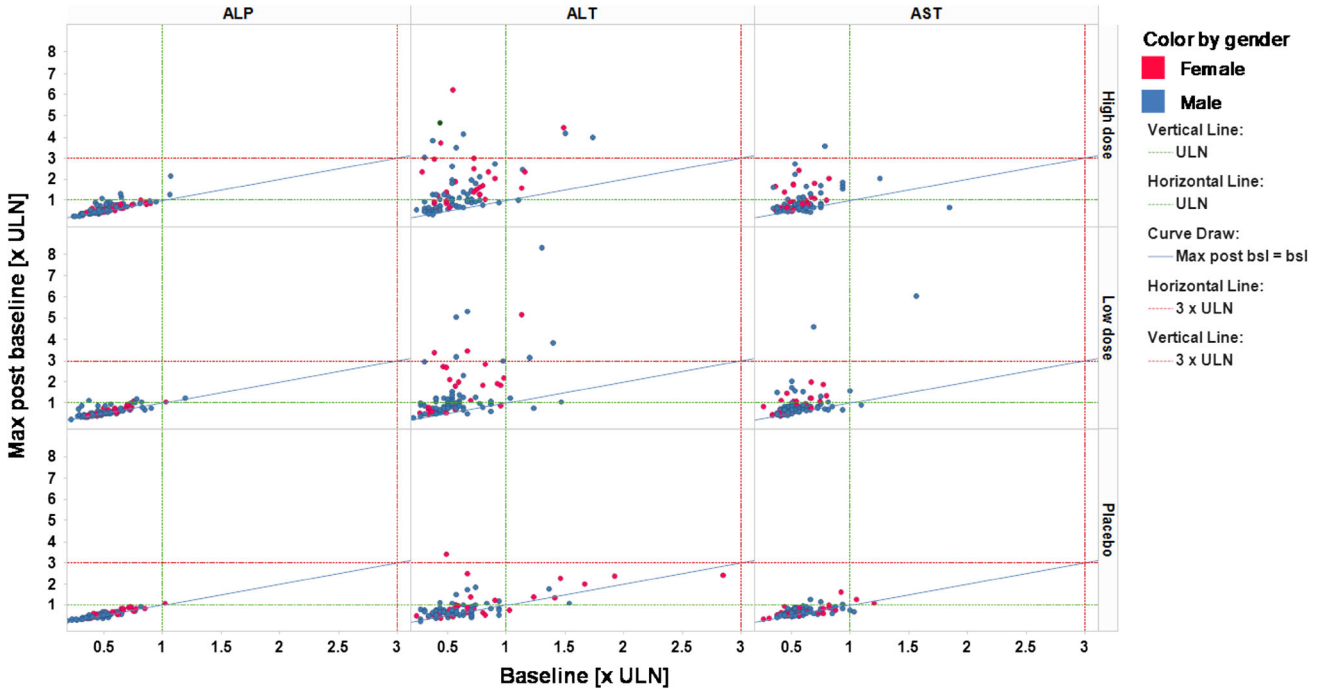

Fig. 5 Shifts from baseline, parameters by column, treatment groups by row, color coding by gender. *ULN* upper limit of normal, *max* maximum, *max post bsl* maximum post baseline

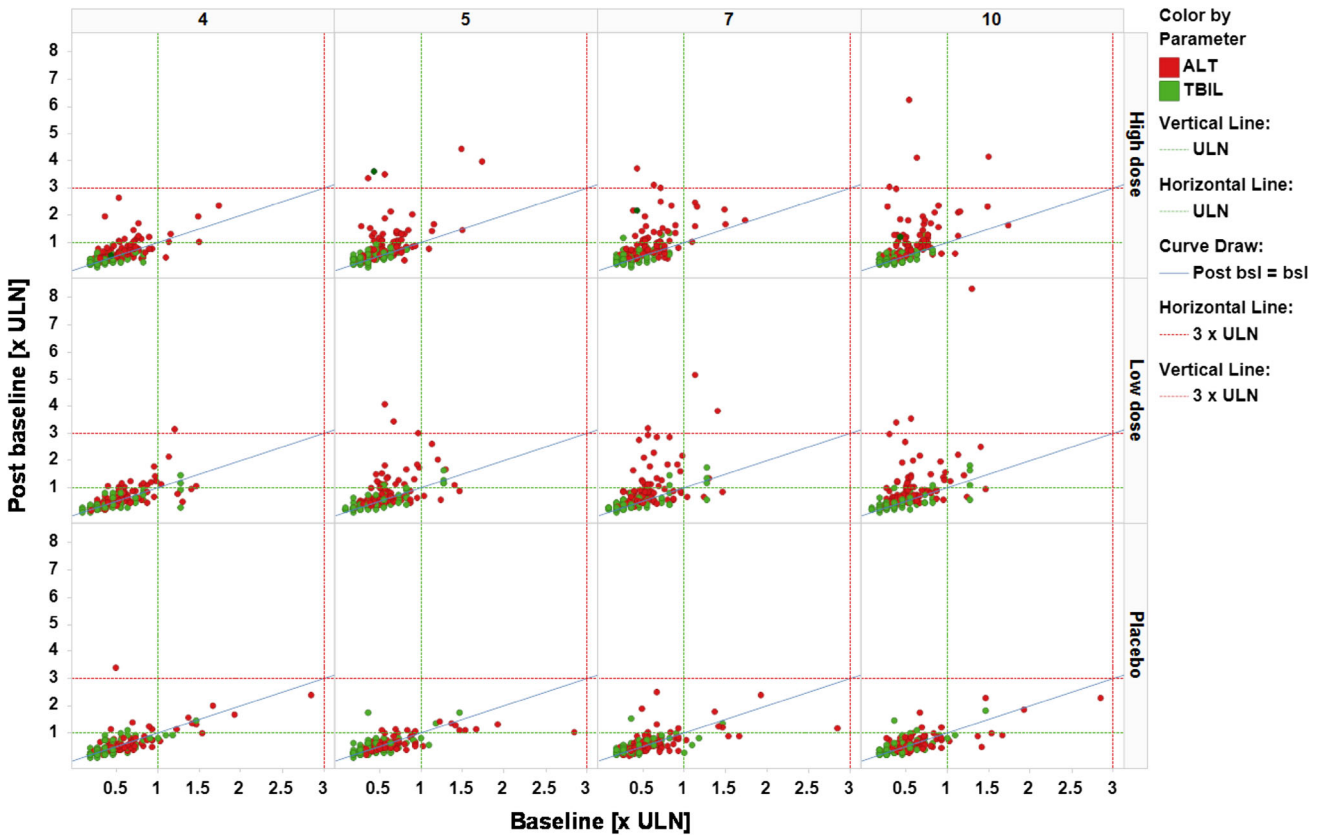

Fig. 6 Shifts from baseline, visits by column, treatment by rows, color coding by parameter. *ALT* alanine aminotransferase, *TBIL* total bilirubin, *ULN* upper limit of normal, *post bsl* post baseline

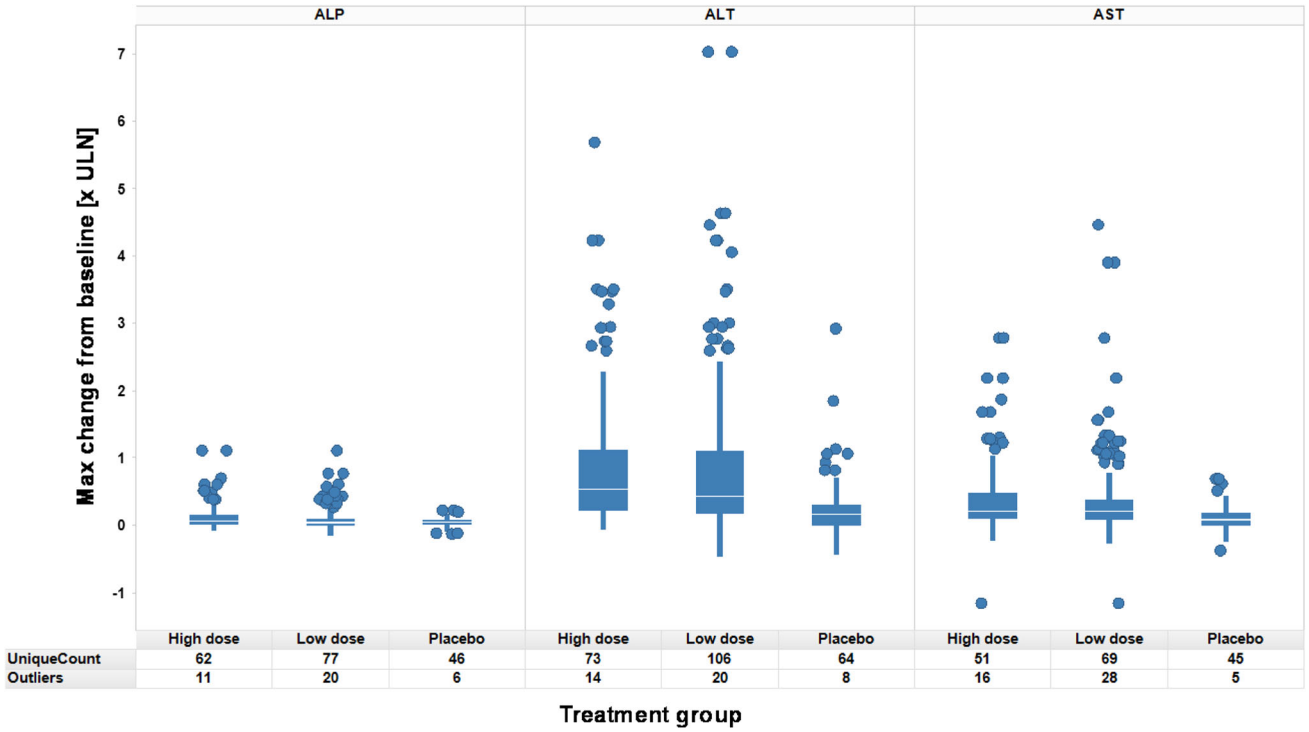

Fig. 7 Maximum absolute changes from baseline across treatment groups, parameters by panel, treatment groups by column per panel

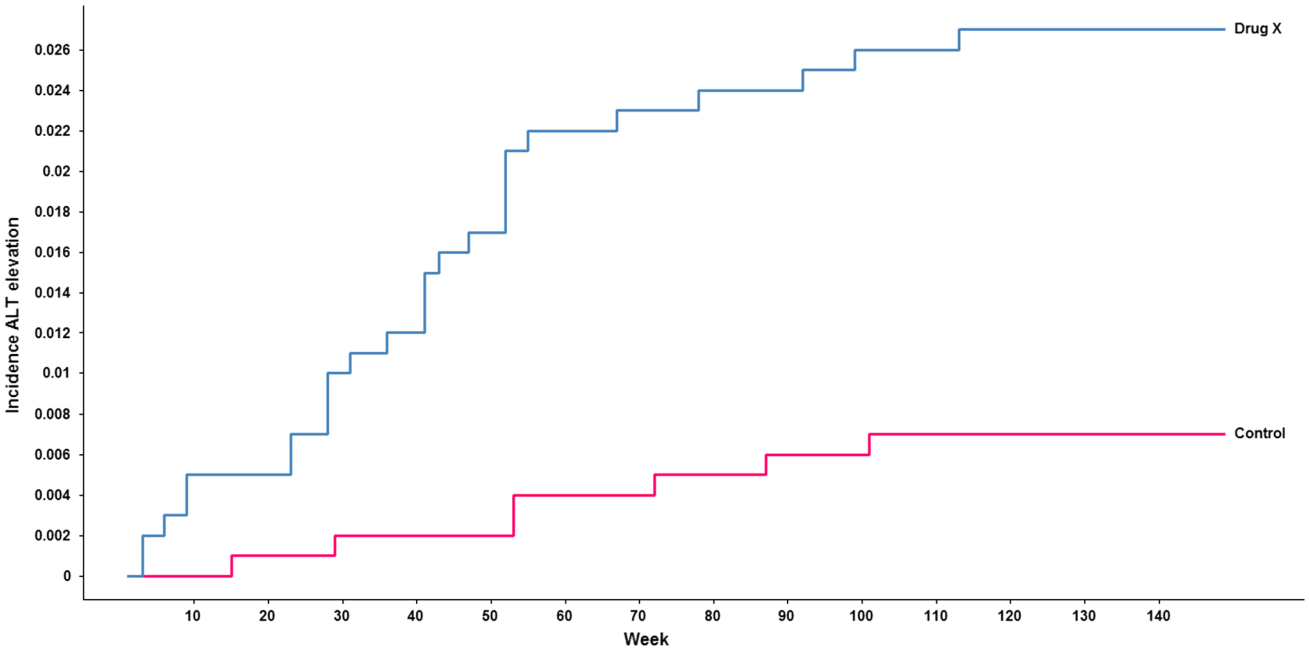

Fig. 8 Kaplan–Meier plot of incidence of ALT elevations over time across treatment groups. *ALT* alanine aminotransferase

the effect may at least partially also be due to a bias introduced by the number of post-baseline measurements, as outlined above.

Figure 6 shows an example displaying shifts by visit, avoiding the bias by multiple measurements when plotting only maximum post-baseline values.

3.2.5 Dose-Response-Relationship

In order to assess dose effects more quantitatively than feasible via scatter plots, box plots may be used for absolute or relative changes from baseline and compared across treatment groups. Figure 7 shows maximum absolute changes from baseline per

patient for liver enzymes across treatment groups. Plots per treatment group are defined by median (white line), lower and upper quartiles (box), lower and upper adjacent values (whiskers), and outliers (individual data points). Outliers are jittered on the x-axis to improve visibility.

The plot suggests differences for maximum elevations from baseline of ALT and AST as compared between both active treatment groups and placebo treatment.

For ALP, only a potential trend towards higher elevations with active treatment as compared to placebo can be observed.

3.2.6 Kaplan–Meier Plots

Capturing and comparing time to elevation of liver test results across treatment groups is of key importance not only for understanding and adequately interpreting a potential liver safety signal but also for managing the risk associated with any effects of the study drug on the liver, e.g. in terms of defining adequate monitoring intervals. The graphical display most widely used to show and compare times to event is the Kaplan–Meier plot, which, in the absence of truncation, censoring, and competing risks corresponds to the empirical cumulative distribution function of incidences to reach a predefined threshold. Such plots can sometimes reveal clear active treatment effects on serum ALT not evident from aggregate data, especially if the active drug treated diseases associated with ALT elevations (e.g. diabetes, congestive heart failure and viral hepatitis).

Figure 8 shows an example of ALT elevations >3 × ULN for drug X (blue line) as compared to control (red line).

4 Other Methods

4.1 Outlier Detection

Particularly for idiosyncratic DILI, identification of abnormal liver chemistry data may be considered as an outlier detection problem. An outlier refers to an observation that deviates markedly from the pattern or distribution of the majority of the data. Graphical displays such as the eDISH plot, shift plots, or box plots, as outlined above, can help substantially to spot clinically relevant outliers in clinical trial data, but sometimes have limited value in terms of reliably differentiate true outliers from random variation. To facilitate that, various robust statistical methods have been proposed, as described in more detail for instance in [15, 17–20]. The following section describes for consideration an approach that has been applied to liver safety data and makes use of both ULN- and baseline-normalized data.

4.1.1 Truncated Robust Multivariate Outlier Detection (TRMOD)

Multivariate outlier detection based on a robust distance measure has been studied extensively and applied to detect outliers in multivariate laboratory data [21]. Mahalanobis distance measures the distance of a subject from the center of the multivariate normal distribution. Multivariate outliers are usually detected based on robust distance which uses the robust estimate of mean and covariance in the calculation of Mahalanobis Distance [17]. The decision boundary for multivariate outlier detection based on a multivariate normal distribution has an ellipsoidal shape in general [19] and is an ellipse for the bivariate (two markers) case (Fig. 9a). The ellipse is a good graphical indicator of the correlation between two variables. The ellipse collapses diagonally as the correlation between the two variables approaches either 1 or −1. The ellipsoid is more circular (less diagonally oriented) if the two variables are less correlated.

Multivariate outliers detected based on such decision boundaries will include outliers in all directions. However, only abnormally high elevations of liver chemistry measurements ALT, AST, ALP, and TBIL indicate a potential

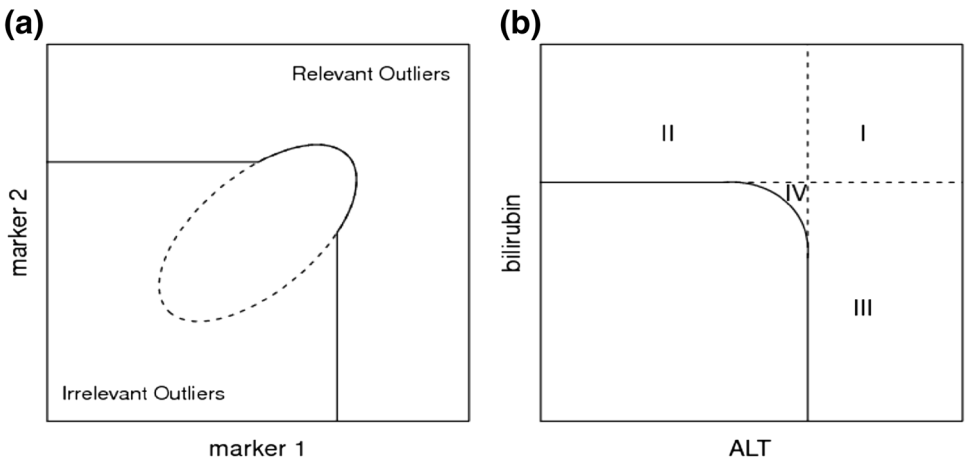

Fig. 9 **a** TRMOD boundary for two correlated measurements. **b** TRMOD boundary for ALT and bilirubin with four regions: (*Region I*) severe toxicity or potential Hy's Law, (*Region II*) elevated bilirubin, (*Region III*) elevated ALT, and (*Region IV*) potentially toxicity. *ALT* alanine aminotransferase

liver safety issue. Hence, outliers with abnormally small values of these liver chemistries are not of interest in identifying potential liver toxicity and would be considered clinically irrelevant outliers. TRMOD [15] was proposed as a robust statistical method for identification of clinically relevant outliers in laboratory safety data while automatically excluding clinically irrelevant outliers. The decision boundary defined by the truncated robust distance is shown in Fig. 9a as a solid line [15]. The TRMOD boundaries are determined based on robust estimates of mean and covariance from the data and can be adjusted based on a specified false detection probability value. In applying TRMOD to liver chemistry data, log transformation of ALT and bilirubin is used so that the majority of the data can be modeled approximately as a multivariate normal distribution. The two liver chemistry measurements, ALT and bilirubin, are not highly correlated, and the decision boundary then appears rounded similar to Fig. 9b.

Truncation lines in TRMOD, i.e. x-limit and y-limit, are determined by their intersection with the horizontal and vertical axis. X-limit and y-limit can be used for comparison with other thresholds based on methods such as eDISH. By extending the truncation line as given in Fig. 9b, using TRMOD it is possible to achieve decision boundaries very similar to eDISH. For the two liver chemistry measurements ALT-limit is interpreted as the x-limit and bilirubin limit as the y-limit. ALT and bilirubin data can be either ULN-corrected or baseline corrected. ALT and bilirubin limits define regions similar to eDISH. Together, they form regions identified as: (Region I) severe toxicity or potential Hy's Law, (Region II) elevated bilirubin, (Region III) elevated ALT, and (Region IV) potential toxicity. Outliers lying in region IV, based on the multivariate analysis, may indicate some abnormality in both ALT and bilirubin simultaneously, requiring further attention. In fact, ignoring region IV, the shape of the decision boundaries are exactly the same in both TRMOD and eDISH. An important difference, however, is that the TRMOD boundaries are estimated from data, in comparison to the eDISH boundaries which are fixed since they were derived from Hy's Law. Since TRMOD based decision boundaries are derived from data and are comparable to eDISH limits for liver chemistry data, it has been suggested to name the associated plot "modified eDISH", "mDISH".

4.1.1.1 Hy's Law Examined by TRMOD In order to compare TRMOD boundaries to empirically derived and fixed thresholds of FDA's Hy's Law limits, data from 28 Phase II–IV clinical trials performed at GSK were aggregated and analyzed by the TRMOD algorithm. ALT and total bilirubin data were analyzed and assessed graphically with the FDA's evaluation of Drug Induced Serious Hepatotoxicity (eDISH) plot assessing fold-ULN, as well as using a modified eDISH (mDISH) plot to assess fold-baseline liver chemistries [22]. The data consisted of 18,672 predominantly female subjects with mean age of 44 years and without known liver disease.

Among generally healthy clinical trial subjects, the empirically derived TRMOD boundaries were

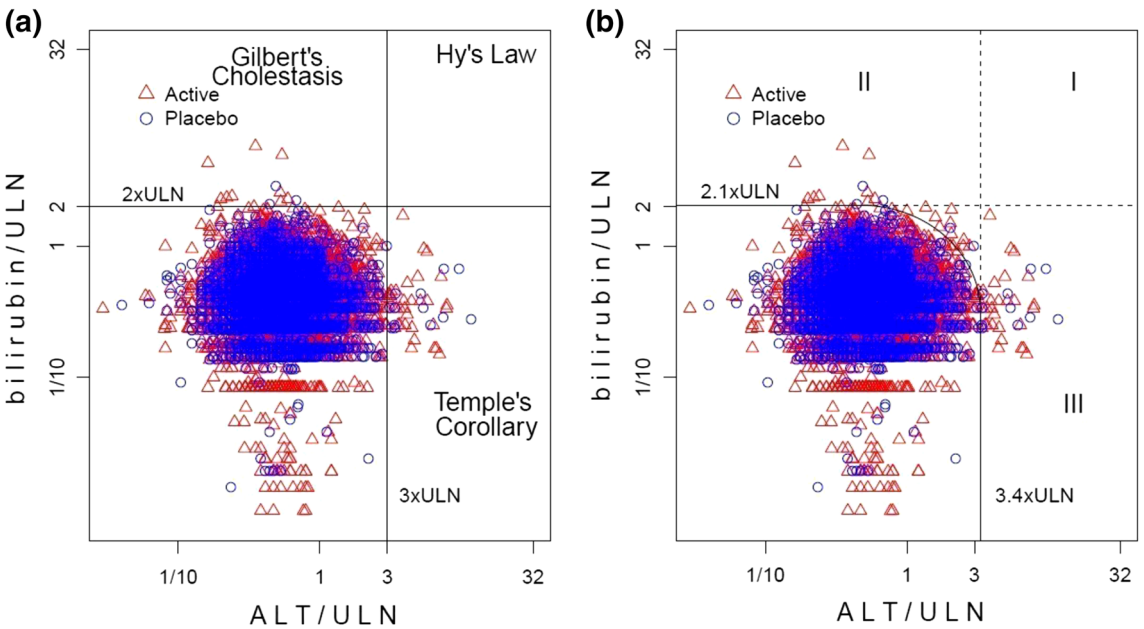

Fig. 10 **a** Hy's Law and eDISH Plot on ULN corrected data: ALT >3 × ULN and total bilirubin >2 × ULN; **b** TRMOD boundaries and mDISH plot on ULN corrected data: ALT >3.4 × ULN and bilirubin >2.1 × ULN. *ALT* alanine aminotransferase, *ULN* upper limits of normal

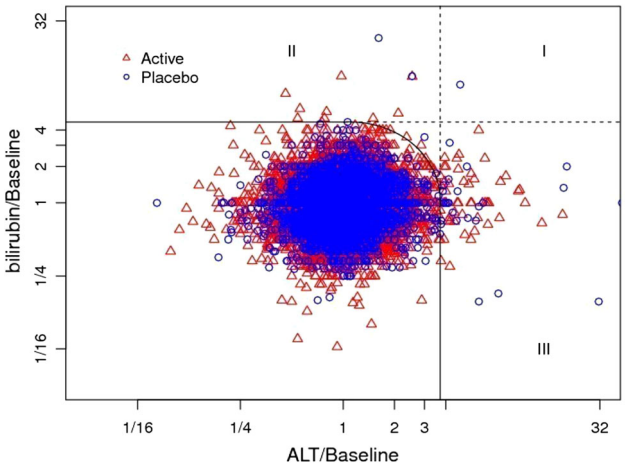

Fig. 11 TRMOD boundaries and mDISH plot on baseline corrected data: ALT >3.8 × baseline and bilirubin >4.8 × baseline. *ALT* alanine aminotransferase

approximately equivalent to 'Hy's Law.' In comparison to FDA's 'Hy's Law' boundaries of 3 × ULN and bilirubin 2 × ULN (Fig. 10a), TRMOD identified outliers with ALT limit of 3.4 × ULN and bilirubin limit of 2.1 × ULN (Fig. 10b). In order to minimize confounding by inter-laboratory variation across the 28 studies, baseline corrected data were used in addition. By applying TRMOD to baseline corrected data, boundaries were 3.8 × baseline for ALT and 4.8 × baseline for bilirubin (Fig. 11).

Overall, TRMOD liver chemistry analyses of clinical trial data in generally healthy subjects confirmed the FDA's Hy's Law threshold as a robust means to detect liver safety outliers. TRMOD evaluation of liver chemistry data, by both fold-ULN and fold-baseline, provides a complementary analysis method and can generate valuable data to establish evidence-based decision boundaries across patient populations. Use of baseline corrected data reduces impact of inter-laboratory variation and may be more sensitive to possible drug effects. However, more data is needed to confirm its value. As long as that data is not available, it is proposed to assess liver chemistries using graphical depictions of both ULN corrected data (eDISH) and modified eDISH (mDISH) for baseline corrected data, as complementary methods.

TRMOD methodology has also been applied to liver chemistry data from 31 aggregated GSK oncology clinical trials to establish population-based thresholds for assessment of liver injury [23]. TRMOD identified outliers with an ALT limit 5.0 × ULN and total bilirubin limit 2.7 × ULN. Additionally, TRMOD was applied to the aggregated oncology data to examine fold-baseline ALT and total bilirubin, indicating outlier detection limits of ALT 6.9 × baseline and bilirubin 6.5 × baseline [23]. Thus, boundaries for outlier detection based on TRMOD methodology were wider in the oncology population as compared to healthy subjects. Similar ALT and bilirubin threshold limits were observed for oncology patients both with and without liver metastases.

4.1.2 *Univariate and Multivariate Extreme Value Modelling*

A slightly different approach to outlier detection has been suggested by Southworth and Heffernan [18, 20] based on extreme value theory. The method estimates probabilities of exceeding thresholds of concern, e.g. ALT >3 × ULN and TBIL >2 × ULN, fitting a generalized Pareto distribution (GPD) to marginal data values above an appropriately chosen threshold, where the effect of baseline on post-treatment values is eliminated by robust regression modelling. Model based probabilities for liver tests exceeding predefined thresholds can support identification of potential liver safety signals even with rather small sample sizes, e.g. predicting incidence and magnitude of outliers in phase III studies or post marketing based on phase II data.

5 Conclusions and Recommendations

Liver safety data from clinical trials usually is complex, multivariate by nature, and typically includes multiple measurements over time. Association of biomarker effects with clinical adverse events, concomitant diseases, and concomitant medication needs to be accounted for.

- In addition to standard summary tables and narratives, graphics can help significantly to detect signals, understand cause–effect relationships, assess mechanisms of toxicity, and may support both risk evaluation and management.
- To facilitate assessment and comparison of liver tests across different parameters, labs, and studies, normalization of data is required. The most common approach of normalizing values by lab-specific upper limits of normal may be supplemented by normalization using the patient's baseline values.
- Proper definition of baseline needs to take into account more than one measurement prior to study treatment. Two measurements at least two weeks apart may be considered a suitable definition.
- A systematic workflow, including predefined graphical templates, starting from the eDISH plot and including a series of line plots, scatter plots, box-plots, and Kaplan–Meier plots helps to ensure completeness of evaluations, supports hypothesis generation and testing, and facilitates identification of the most suitable graphics for publishing and regulatory reporting.

- In particular for detection and assessment of idiosyncratic DILI, attention needs to be given to outliers in a dataset rather than just mean and median values of liver tests. To differentiate true outliers from random variation, robust statistical methods are available that should be considered for liver safety evaluation.

Acknowledgements The Innovative Medicines Initiative and the Hamner—University of North Carolina Institute for Drug Safety Sciences sponsored the workshop, part of which is summarized in this article. This article is part of a supplement entitled *Liver Safety Assessment in Clinical Drug Development: A Best Practices Workshop report*, which was guest edited by Drs. Paul B. Watkins, Michael Merz and Mark I. Avigan. The guest editing by Dr. Avigan does not reflect the position of, nor imply endorsement from, the US Food and Drug Administration or the US Government. Drs. Watkins, Merz and Avigan did not receive any honoraria for guest editing the supplement. All manuscripts were peer reviewed by Dr. Rolf Teschke. Dr. Rolf Teschke has no conflicts of interest to declare and did not receive any honoraria for peer reviewing the supplement; however, he received a free yearly online subscription to the journal *Drug Safety*.

The Innovative Medicines Initiative (http://www.imi.europa.eu/) is a public-private partnership set up by the European Commission in 2008 to relieve the bottlenecks in drug development and to provide economic stimulus. With a €2 billion commitment, the IMI now has an important portfolio of projects where experts from academia, industry and regulatory bodies collaborate on an unprecedented scale and at a non-competitive level to develop tools and technologies. Drug-induced liver injury has been a focus of several projects including the SAFE-T (Safer and Faster Evidence-based Translation) consortium, which is working on clinical qualification of new biomarkers to better detect and characterize liver toxicity, and MIP-DILI, which is working to determine the optimal preclinical testing to detect potential of liver injury in patients.

The Hamner-University of North Carolina Institute for Drug Safety Sciences (IDSS—http://www.thehamner.org/idss/), located in Research Triangle Park, NC, is dedicated to solving drug safety challenges through a variety of innovative approaches including mouse genetics, mechanistic biomarkers, and culture models derived from induced pluripotent stem cells. Efforts in drug-induced liver injury include the DILI-sim Initiative, a public-private partnership developing computer models to explain and predict drug-induced liver injury.

Conflict of interest Michael Merz, Kwan Lee, Gerd Kullak-Ublick, Andreas Brueckner, and Paul Watkins have no conflicts of interest that are directly relevant to the content of this article.

References

1. US Department of Health and Human Services FDA, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). Guidance for industry. Drug-induced liver injury: premarketing clinical evaluation. 2009;18.
2. Ozer J, Ratner M, Shaw M, Bailey W, Schomaker S. The current state of serum biomarkers of hepatotoxicity. Toxicology. 2008;245(3):194–205.
3. Shi Q, Hong H, Senior J, Tong W. Biomarkers for drug-induced liver injury. Expert Rev Gastroenterol Hepatol. 2010;4(2):225–34.
4. Whitfield J. Gamma glutamyl transferase. Crit Rev Clin Lab Sci. 2001;38(4):263–355.
5. Krause A, O'Connell M. A picture is worth a thousand tables: graphics in life sciences. New York: Springer; 2012.
6. Cleveland W. The elements of graphing data. Summit: Hobart Press; 1994.
7. Cleveland W. Visualizing data. Summit: Hobart Press; 1993.
8. Tukey J. Exploratory data analysis. Essex: Addison-Wesley; 1977.
9. Weil JG, Bains C, Linke A, Clark DW, Stirnadel HA, Hunt CM. Background incidence of liver chemistry abnormalities in a clinical trial population without underlying liver disease. Regulat Toxicol Pharmacol. 2008;52(2):85–8.
10. Cai Z, Christianson AM, Stahle L, Keisu M. Reexamining transaminase elevation in phase I clinical trials: the importance of baseline and change from baseline. Eur J Clin Pharmacol. 2009;65(10):1025–35.
11. Dutta A, Saha C, Johnson CS, Chalasani N. Variability in the upper limit of normal for serum alanine aminotransferase levels: a statewide study. Hepatology. 2009;50(6):1957–62.
12. Sibille M, Deigat N, Durieu I, Guillaumont M, Morel D, Bienvenu J, Massignon D, Vital Durand D. Laboratory data in healthy volunteers: reference values, reference changes, screening and laboratory adverse event limits in phase I clinical trials. Eur J Clin Pharmacol. 1999;55:13–9.
13. Crowe B. Core analyses for program-level safety reviews (PLSRs). Chicago: DIA meeting; 2011.
14. Gelperin K, Guo T, Senior J. A simple graphic tool for assessing serious liver injury cases in a clinical trial—eDISH. 2008 (online). http://www.fda.gov/drugs/scienceresearch/researchareas/ucm076901.htm. Accessed 21 October 2013.
15. Lin X, Parks D, Zhu L, Curtis L, Steel H, Rut A, et al. Truncated robust distance for clinical laboratory safety data monitoring and assessment. J Biopharm Stat. 2012;22(6):1174–92.
16. Puukka K, Hietala J, Koivisto H, Anttila P, Bloigu R, Niemela O. Additive effects of moderate drinking and obesity on serum gamma-glutamyl transferase activity. Am J Clin Nutr. 2006;83(6):1351–4.
17. Trost DC. Multivariate probability-based detection of drug-induced hepatic signals. Toxicol Rev. 2006;25(1):37–54.
18. Southworth H, Heffernan JE. Extreme value modelling of laboratory safety data from clinical studies. Pharm Stat. 2012;11(5):361–6.
19. Maesschalck R, Jouan Rimbaud D, Massart DL. The Mahalanobis distance. Chemom Intell Lab Syst. 2000;50:1–18.
20. Southworth H, Heffernan JE. Multivariate extreme value modelling of laboratory safety data from clinical studies. Pharm Stat. 2012;11(5):367–72.
21. Southworth H. Detecting outliers in multivariate laboratory data. J Biopharm Stat. 2008;18(6):1178–83.
22. Lin X, Parks D, Painter J, Hunt CM, Stirnadel-Farrant HA, Cheng J, et al. Validation of multivariate outlier detection analyses used to identify potential drug-induced liver injury in clinical trial populations. Drug Saf. 2012;35(10):865–75.
23. Parks D, Lin X, Painter JL, Cheng J, Hunt CM, Spraggs CF et al. A proposed modification to Hy's law and Edish criteria in oncology clinical trials using aggregated historical data. Pharmacoepidemiol Drug Saf. 2013 (epub ahead of print).