

Evolution of the Food and Drug Administration Approach to Liver Safety Assessment for New Drugs: Current Status and Challenges

John R. Senior

© The Author(s) 2014. This article is published with open access at Springerlink.com

Abstract Prompted by approval in 1997 of troglitazone and bromfenac, two drugs that promptly began to show serious and sometimes fatal liver toxicity, we began at the Food and Drug Administration (FDA) a series of annual conferences in 1999 to consider issues of drug-induced liver injury (DILI). First inviting reviewers of new drug applications we opened the audiences in 2001 to pharmaceutical industry and academic consultants to industry and FDA, and slides shown at the meetings were posted on the internet to be available at the website of the American Association for the Study of Liver Diseases (AASLD)—go to (<http://www.aasld.org/dili/Pages/default.aspx>). Observations by Dr. Hyman J. Zimmerman that “drug-induced hepatocellular jaundice is a serious lesion” with possible mortality formed a basis for developing a computer program to plot peak serum values for alanine aminotransferase (ALT) and total bilirubin (TBL) in an x–y log–log graph for all subjects enrolled in clinical trials. This program had the capability to show the time course of all liver tests for individuals who had both hepatocellular injury and reduced whole liver function, plus clinical narratives to diagnose the severity and most likely cause of the abnormalities. We called the program eDISH (for evaluation of **D**rug-**I**nduced **S**erious **H**epatotoxicity), and began in 2004 to use it to assess DILI in clinical trial subjects. From 2008,

comments made by the presenters at the conferences about their slides and ensuing discussions have been added to the website. All this has raised awareness of the problem, and since 1997, the FDA has not had to withdraw a single drug because of post-marketing hepatotoxicity. Many issues still remain to be resolved; among the most controversial is the best method to estimate likelihood that a given liver injury was actually caused by the drug in question. On November 9, 2012, a workshop was convened to discuss the best practices for the assessment of drug-induced liver injury (DILI) in clinical trials.

Key Points

Severity of liver injury cannot be determined by ALT elevations alone, and receiver-operating characteristic values often fail for very rare events

DILI cannot be diagnosed by serum chemistries alone, nor by liver biopsy, but requires pertinent clinical information

A new biomarker will have to be extremely specific to be useful, and a better method is needed to estimate if a given liver injury was caused by the drug

The views expressed in this article are those of the author and are not a summary of the official position and policy of the FDA.

J. R. Senior (✉)
Office of Pharmacovigilance and Epidemiology,
Office of Surveillance and Epidemiology, Center for Drug
Evaluation and Research, Food and Drug Administration,
10903 New Hampshire Avenue, Silver Spring,
MD 20993-0002, USA
e-mail: john.senior@fda.hhs.gov

1 Introduction

Timely detection and proper assessment of drug-induced liver injury (DILI) in clinical trials has been for decades 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 and The Hamner Institute for Drug Safety Sciences, with the aim of addressing gaps in current guidance and initiating alignment of liver safety assessment on a global scale.

On November 9, 2012, regulatory experts from the FDA, European Medicines Agency, Health Canada, and the Japanese National Institute of Health Sciences, with representatives from industry and academia, convened in Boston and discussed what could be considered best practices in clinical liver safety assessment, focusing 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 the Evolution of the Food and Drug Administration Approach to Liver Safety Assessment for New Drugs: the current status and challenges. This brief historical note is written from the perspective of an academic gastroenterologist who has had a special interest in liver disorders (25 years), plus pharmaceutical industry experience in work on gaining approval for new drugs (16 years), and who has been employed (19 years) as a gastrointestinal medical reviewer and consultant in hepatology at the Food and Drug Administration (FDA). It is not intended to be a comprehensive review of the subject of drug-induced liver injury, nor an attempt to resolve still-controversial issues in the field.

2 Current Status

2.1 Safety Withdrawals for New Molecular Entities Approved by FDA in 1997

In 1997, several drugs were approved by the FDA that generated criticism [1] because they appeared to cause serious, sometimes fatal injury to vital organs such as the liver, heart and skeletal muscle (Table 1).

Of special hepatological interest were two new drugs, troglitazone and bromfenac, approved in January and July 1997. Troglitazone [2] was the first thiazolidinedione for treatment of type-2 diabetes mellitus; bromfenac [3] was just another non-steroidal analgesic drug for pain relief. Both of these new drugs rather promptly began to cause serious liver injury, liver failure, and death in patients [4, 5]. Bromfenac was withdrawn from the US market in June 1998; troglitazone had been taken off the market in the UK in 1997 [6] but not until May 2000 in the US [7] after two alternative thiazolidinedione agents, rosiglitazone and pioglitazone, were approved.

2.2 Office of Post-Marketing Drug Risk Assessment (OPDRA)

The Center for Drug Evaluation and Research established in 1998 an Office of Post-Marketing Drug Risk Assessment (OPDRA), and published a response to critical press comments [8]. The new OPDRA, focused on drug safety, was intended to evaluate reports of adverse drug effects after marketing of new drugs, when greater numbers of patients were treated than had been studied in clinical trials. Spontaneous, voluntary reports to the Adverse Events Reporting System (AERS) had been usually made first by prescribing physicians reporting to the company sponsor of the drug that an adverse effect had been noted in a patient receiving it, after which reports were forwarded by the sponsor to the Food and Drug Administration. Large numbers of reports accumulated and required development of new statistical analyses for observed associations [9]. However, the information content in most of the AERS reports was insufficient to determine accurately either the clinical severity of the adverse effect or its most likely cause.

2.3 FDA Annual Educational Conferences on Drug-induced Liver Injury

Stimulated by adverse liver effects of these new drugs, we proposed in the summer of 1998 that an educational course

Table 1 Safety withdrawals for new molecular entities approved by FDA in 1997

| New drug | Date approved | Date withdrawn | Reason for withdrawal |
|----------------|---------------|----------------|---------------------------------|
| Troglitazone | 29 Jan 1997 | 21 Mar 2000 | Liver toxicity |
| Mibefradil | 20 Jun 1997 | 8 Jun 1998 | Fatal arrhythmias |
| Cerivastatin | 26 Jun 1997 | 8 Aug 2001 | Rhabdomyolysis, renal failure |
| Bromfenac | 15 Jul 1997 | 22 Jun 1998 | Severe hepatitis, liver failure |
| Sibutramine | 22 Nov 1997 | 8 Oct 2010 | Risk of heart attack, stroke |
| Trovafloxacin | 18 Dec 1997 | 16 Jun 2006 | Liver toxicity |
| Alatrofloxacin | 18 Dec 1997 | 16 Jun 2006 | Liver toxicity |

be prepared for FDA reviewers on the interplay of the liver functions and drug effects, called “Drugs and the Liver: What They Do to Each Other.” A course outline was developed, and we sought advice about it from an esteemed consultant in the emerging field of drug-induced hepatotoxicity, Dr. Hyman J. Zimmerman. He supported and encouraged it and offered to speak. We obtained permission to offer the course, announcing it in September 1998. In April 1999, about 325 FDA reviewers attended the two-day course at the University of Maryland campus at Shady Grove Conference Center, with support and participation by CDER leadership. Dr. Zimmerman developed health problems and asked Dr. James Lewis to present his material, but he was able to attend the conference, his last public appearance before his death in July 1999. An eponymic appellation [10], “Hy’s Law,” was put forth there by Dr. Robert Temple of the FDA, to which Dr. Zimmerman objected, but he did agree with the principle. First stated in his 1968 Kober Lecture [11] at Georgetown University, his observation was repeated in both editions of his texts of 1978 and 1999: “**drug-induced hepatocellular jaundice is a serious lesion,**” with mortality ranging from 10 to 50 %. A short course for 75 more reviewers was given in November 1999. This educational approach has been continued with a series of annual conferences on drug-induced liver injury (DILI), the 14th of which was held in March 2014 (slides and verbatim text of comments available on the internet <http://www.AASLD.org> under the header Training/Education).

In January 2000, the author began working in the new OPDRA, devoting full time to responding to consultation requests from the medical review divisions of the Office of New Drugs on cases of possible DILI detected in clinical trials of drugs under review for approval. In 2001, OPDRA was renamed the Office of Drug Safety, in 2003, renamed the Office of Pharmacoepidemiology and Statistical Science (OPaSS), and in 2006, the Office of Surveillance and Epidemiology (OSE). Over the 13 years since, scores of requests for consultation about possible drug-induced liver toxicity have been received and answered. These consultation requests have been difficult, challenging, nearly always concerning a new drug compound never before heard of, for treating a disorder of which little or nothing was known by the author, asking for a learned reply rapidly. Medical literature, Google searches, discussions with experts, whatever was needed, were used to understand the problem, obtain and examine the data available, then give explanations and recommendations. These consultation responses were confidential, intended for the review division staff and other FDA/CDER staff only. They were not published openly, and were not circulated outside the Agency. The learning process led to recording, sharing, and teaching

what had been learned, and to seeking additional ideas and reactions from others, forming the basis for the annual public DILI conferences mentioned above.

2.4 Causality of the Observed Effect

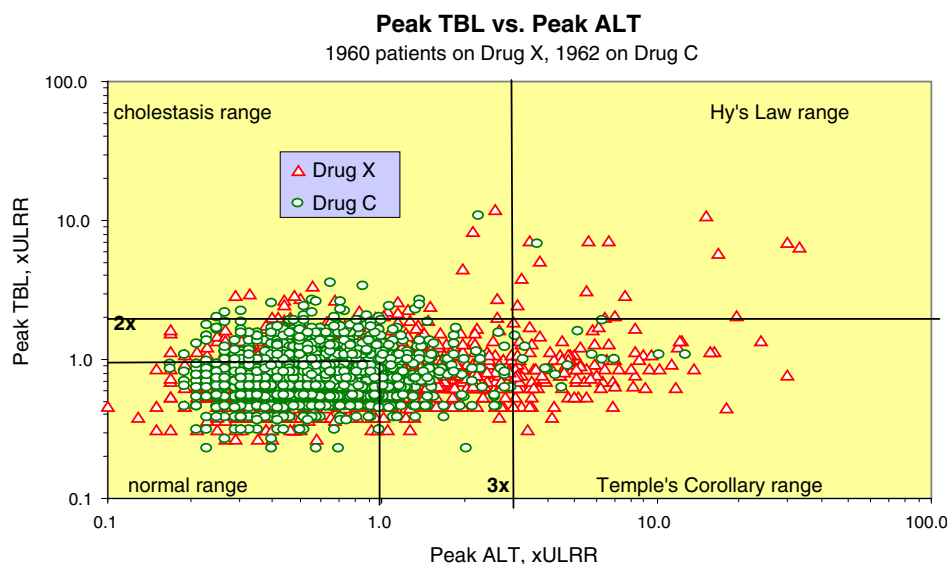
Primary among the key elements of the Zimmerman dictum was the concept of *causality* of the observed effect as drug-induced, rather than caused by viral infection, alcohol or other drug overdose, autoimmune effects, or many alternative possible causes. It was recognized that DILI could mimic almost any other known form of liver disease, could not be diagnosed accurately even by liver biopsy and histologic interpretation, had no biomarker that could be relied upon confidently, and serious DILI was a rare occurrence from approved drugs. The diagnosis of DILI became a diagnosis of exclusion, arrived at by a series of studies and interrogations to exclude other possible causes of the findings. This process of medical differential diagnosis is an art for which physicians are trained and become experienced and skillful. The difficulty of determining the probable cause of abnormal liver test findings of “drug-induced hepatic injury” was pointed out by Popper and Schaffner [12] a decade earlier than the Zimmerman observation.

2.5 eDISH

The Hy’s Law principle then formed the basis for the FDA’s development of a software program called “eDISH” (for evaluation of **drug-induced serious hepatotoxicity**). The program was written in a language called SASIntrNet by an insightful and open-minded expert statistician, Dr. Ted Guo. It allowed scanning of large numbers of patients from clinical trials for indicators of hepatocellular injury by serum alanine aminotransferase (ALT) activity and of whole liver dysfunction by total bilirubin (TBL) concentration. These biomarkers were routinely measured in almost all clinical trials, usually with serial values obtained according to protocol. For the first step in the eDISH analysis, data from case reports for all patients or subjects enrolled in a trial were surveyed for peak values of ALT and TBL over their entire period of observation. See Fig. 1.

They were plotted on an x–y chart as \log_{10} values of multiples of elevations above the upper limits of the normal reference ranges (\times ULN), to bring the much greater fluctuations of ALT than TBL into a corresponding scale, yet preserve their magnitudes. For conservative purposes, to preserve sensitivity of detecting nearly all cases, low cut-off levels were employed: $3\times$ ULN for ALT and $2\times$ ULN for TBL. Using data from all of the subjects or patients in a clinical trial, peak values for ALT and TBL

Fig. 1 A step-one eDISH plot of almost 4,000 randomized patients. *ALT* alanine aminotransferase, *TBL* total bilirubin, *eDISH* evaluation of drug-induced serious hepatotoxicity, *ULRR* upper limit of reference range or normal



are plotted as a planar x–y graph, with ALT abscissa values and ordinate TBL values, each point representing a single person.

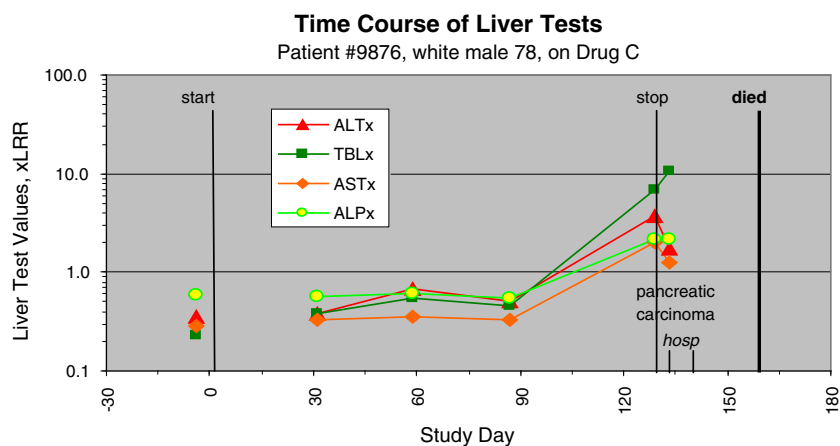
Shown in Fig. 1, is an eDISH plot for almost 4,000 patients enrolled in a trial of an experimental new drug (X) and a control agent (C), 1960 randomized to X, 1962 to C. The cut-off lines then make four quadrants: a lower left quadrant contains most patients with normal or near normal peak values for both variables; an upper left quadrant that contains those with elevated bilirubin levels but not much ALT elevation; a lower right quadrant showing those with elevated ALTs but not TBLs, indicating hepatocellular injury without whole liver dysfunction; and a few patients in the right upper quadrant who showed clinically significant hepatocellular injury AND whole liver dysfunction by elevations of both, not necessarily on the same test day. Cut-off levels were not established by data analyses, but were based on expert opinions at a Fogarty International Conference in 1978 that concluded [13] that ALT values $>3 \times \text{ULN}$ and TBL $>2 \times \text{ULN}$ were “markedly abnormal.” No adjustment was made for the general practice that each local laboratory establishes its own ranges of normal values, with variance in such reference ranges. *The cutoff lines do not mean that those in the upper right quadrant can be called “Hy’s Law” cases, but only that they are identified as patients of special interest for which more clinical information should be sought for medical diagnosis of the most likely cause of the findings.* The usefulness of the eDISH program has been appreciated by many pharmaceutical companies and widely copied, at least for the step-one x–y log–log plot of ALT and TBL, especially with the encouragement of some of their principal consultants.[14]

The second step of the eDISH program (Fig. 2) is activated by pointing to and clicking on a symbol on the first x–y plot, which brings up the time course of ALT,

TBL, plus aspartate aminotransferase (AST) and alkaline phosphatase (ALP) for that single patient over the time of observation, all four variables plotted together for easy visual comparison. It can be noted at a glance if the bilirubin elevation preceded or followed the rise in ALT or AST, and when there were notable changes in serum chemistry test values. In the case selected above (the single green circle on the first x–y plot indicating randomization to control drug, warfarin), the time course and clinical narrative revealed that the patient had undiagnosed pancreatic carcinoma, very unlikely to have been caused by warfarin, and fatal to the patient a short time later. In responding to many consultation requests, we have found it useful to request data from the sponsor on all liver tests done during clinical trials, including those from local laboratories, for all subjects or patients in selected and pivotal clinical studies, in eDISH format as specified by Dr. Guo. Many companies have copied the first eDISH plot with programs for their own use.

For diagnosis of DILI or some other cause, the time course is greatly augmented by a medical text narrative about that person, as eDISH step three. They should be written by a physician skilled in medical differential diagnosis, to estimate the most likely *cause* of the abnormal findings and clinical picture, and to make an assessment of the probability that it was drug-induced or not. It is seldom possible to be absolutely certain that the findings were caused by the drug and not by something else. That is still a medical art. If there are several possible causes, and there usually are, it is not enough simply to conclude that a case is thought to be “confounded;” they nearly all are. Protocols do not always anticipate adverse effects; simple case report forms may not be useful in diagnosing causality. We take an estimated likelihood as “probable” if $>50\%$ likely, more than all other causes combined, If

Fig. 2 The time course of liver tests



information provided is sufficient, and the patient has been well investigated to rule out alternative causal possibilities. We may be able to raise the likelihood level to >75 %, “very likely,” or >95 %, “almost certain,” with additional medical information. For less well supported diagnoses, we have taken >25 to 50 % as “possible,” 5 to 25 % as “unlikely,” and <5 % as “very unlikely.” Admittedly, these are not exact determinations but are clinical estimates based on the skill and experience of the rater and quality of the information provided.

It is clear that correct diagnosis cannot be made by serum chemistry abnormalities alone. The use of the term “biochemical Hy’s Law” reflects profound misunderstanding of the purpose and use of eDISH, but is a term all too commonly mis-used by company statisticians for screening new drugs in development. We avoid terms saying adverse findings were “associated with” or even “related to” the drug, but mean that the term “drug-induced” indicates at least *probably caused by* the drug, not by something else.

Is this distinction important? It certainly is, for what we are seeking to find is usually rare, on the order of 1 per 1,000 to 10,000 who are exposed to the drug suspected. For rare events, test sensitivity is far less important than high specificity. A test or biomarker to detect abnormal states is very dependent upon the prevalence or incidence of the abnormality sought. We seek to find, as early as possible before drug-induced injury is irreversible, *serious* DILI that causes disability, requires hospitalization, and leads to acute liver failure, death or liver transplantation. We are less concerned with serum aminotransferase elevations that are transient, asymptomatic, not accompanied or followed by degradation in liver functions such as clearance of bilirubin or synthesis of prothrombin. The liver is a remarkable organ, capable of regeneration even when 65 % of it is destroyed or resected, able to adapt and become tolerant of xenobiotic agents, and of variable genetic ability to tolerate new drugs. Only rare persons are unable to tolerate or to

adapt to the new drug. This is not a statistical but a medical problem; the statistical “outlier” is the medical patient of interest who needs investigation to find out what is causing the abnormal findings before it is too late.

This concept of setting grades for elevations of serum enzyme activities was established in 1983 when the National Cancer Institute (NCI) of the United States National Institutes of Health (NIH) first published their Common Toxicity Criteria (CTC) for adverse events. Results of laboratory chemistry tests, physical findings, and indicators were included, based on opinions of experts or committees. The CTC assessed serum enzyme elevations as grade 1, mild (>ULN to 2.5×ULN); grade 2, moderate (>2.5×ULN to 5×ULN); grade 3, severe (>5×ULN to 20×ULN); and grade 4, life-threatening (>20×ULN). These have been updated periodically; version 3 changed the name in 2006 to the Common Terminology Criteria for Adverse Events (CTCAE), and version 4 in 2009, modified the cut-off between mild and moderate ALT and AST elevations from 2.5 to 3×ULN. The CTCAE are widely used and followed by oncologists and other disciplines [15].

Test results are usually interpreted as either “positive” or negative. Determination of what may be the optimal cut-off value is critical. High sensitivity to detect the disease by positive results and not miss many is important, as is high specificity so that normal people without disease show negative test results. In the real world, no test is perfect. The trade-off must be endured: more sensitivity for less specificity, or the reverse. This idea was captured brilliantly by the popular receiver-operating characteristic (ROC) curves emerging from World War II detection by radar signals of enemy planes approaching, then developed for medical use [16] in psychology, then radiology, and recently more generally in medicine as a convincing way to evaluate test results or “signals.” For ROC curves, test sensitivity is plotted on the ordinate and unspecificity (1 – specificity) on the abscissa, so what is being examined is

the true positive versus false positive test results. Such ROC curves have become a widely used to “validate” new tests, but too often this neglects a critical third element: incidence, or prevalence, which must be appreciated. Let us assume that a new test or biomarker has perfect 100 % sensitivity, detecting all cases being sought and missing none, with impressive specificity of 95 %, showing only uncommon false positive results, but applied to detecting a relatively rare (1 per 1,000) problem. For new-onset liver injury caused by a drug, usually a relatively rare event, incidence of 1/1,000 is not uncommon, and serious cases are often even less frequently seen. Because the great majority, almost all, people do not show DILI, the value of positive results is low (see Table 2).

If we use our new test on 100,000 persons to detect a disorder with an incidence of 1 per 1,000, please note the requirement for high specificity to avoid overwhelming the true positive results with false positives. It is obvious from this that even a test or biomarker with perfect sensitivity and high specificity allows correct interpretations from positive test results in less than 2 % of them, so that more than 98 % are wrong, if looking for a rare problem! Pursuit of diagnoses by investigative work-ups in all the >98 % of patients tested with false positive results has raised costs of assessing approved drugs, and has also defeated routine monitoring attempts for many years. Both patients and physicians grow weary of repeatedly negative test results, and soon quit trying. To be >90 % correct would require test specificity of 99.99 %. Such specifications for biomarkers are unheard of, and it is extremely difficult to show such values of sensitivity and specificity for candidate biomarkers proposed as “better” than the combination of {ALT>3×ULN & TBL>2×ULN}. Time-course and narrative data, as used in the eDISH approach, sharpen differential diagnosis even more, and provide the very high specificity needed.

The “values” of positive and negative test results [17] are often referred to as “positive predictive values, PPV,” or “negative predictive values, NPV,” and NPVs are

Table 2 Key importance of specificity when screening for rare events

| Test result | DILI | No DILI | Totals | Value |
|-------------|-------------|-------------|----------|--------|
| Positive | 100 (TP) | 4,995 (FP) | 5,095 | 0.0196 |
| Negative | 0 (FN) | 94,905 (TN) | 94,005 | 1.0000 |
| | 100 | 99,900 | 100,000 | |
| Incidence | Sensitivity | Specificity | Accuracy | |
| 1 per 1,000 | 100 % | 95 % | 95.0 % | |

DILI drug-induced liver injury, the disease sought, *TP* true positive test result, *FN* false negative test result, *FP* false positive test result, *TN* true negative test result

sometimes cited as evidence of how good a test may be. This is grossly erroneous, and is an example of misunderstanding that still needs to be corrected. The values may indicate what *has happened* to date, but not what *will happen*. However, use of serial values to establish trends over time may be helpful. Rather than referring to the tests results as “predictive,” it would be preferable to call them indicative or diagnostic.

Whether the educational efforts and national conferences have raised awareness about DILI problems, or whether increasing use of eDISH approach by sponsors and reviewers have had an impact, the FDA has not approved any drug since the end of 1997 that later had to be taken off the market because of serious hepatotoxicity. This should be viewed from the perspective [18] of the 740 new drugs approved by FDA in the three decades from 1980 through 2009 of which 118 (16 %) had been withdrawn as of December 2010, but only 31/118 (26 %) for safety reasons. As noted, 1997 was an especially bad year for drug approvals later withdrawn for safety problems, four because of liver toxicity, two for cardiac, and one for skeletal muscle injury (Table 1).

“Withdrawals” may be affected by FDA rescinding of approvals, but more often are done by sponsors who voluntarily stop marketing the drug because of adverse reports, but may vary widely between market availability and regulatory action. The seven drugs in Table 1 were withdrawn after varying delays. Trovafloxacin hepatotoxicity was first noted in 1998 but not published [19, 20] until 2000, and it was not officially withdrawn until 2006, so was not included in the initial course for reviewers.

The present approach of responses to consultations going directly to the review divisions considering new drugs for approval, with the Office of Surveillance and Epidemiology also being notified, seems to have been effective in contributing to the prevention of serious DILI. It may supplement the larger efforts on data mining and screening of large numbers of voluntary, spontaneous reports to the AERS system for detecting adverse effects of marketed drugs after they have occurred. It seems better to *prevent* approval of new drugs that show indications of possibly serious liver toxicity in clinical trial data, than to discover it later after approval and marketing. The quality of clinical data for assessing severity and, more importantly, *probable causality*, of adverse liver effects requires the more complete data from clinical trials.

3 Current Challenges in Assessing Liver Safety of Drugs: What is the Way Forward?

It is appropriate to ask what current challenges exist and how this record may be improved. We do not yet have

biomarkers that are specific enough to surpass medical differential diagnosis. If a liver injury is probably drug-induced, affects enough hepatocytes that the whole organ is not able to clear plasma of bilirubin so jaundice occurs, and not mainly cholestatic, Temple found that the Zimmerman rule was just about always correct, that is, very highly specific, *if probably caused by the drug*. A challenge remains that expert hepatological judgment is often needed to diagnose DILI convincingly, with considerable variance in the judgments, even among experts. The several hepatologists of the NIH-supported drug-induced liver injury network (DILIN) have struggled with this for eight years but have no better alternative, and still cannot agree [21] upon the reliability of the RUCAM (Roussel-Uclaf Causality Assessment Method), an algorithmic process [22, 23] proposed in 1993 and used widely in Europe and elsewhere in the world [24]. It is also recognized that exact diagnosis of causality cannot be made by liver biopsy, although useful information about the state of the liver can be learned [25].

It was also observed by Temple that drugs which caused a marked preponderance of lesser liver injuries compared to control agents, shown by more frequent serum aminotransferases, were more likely also to show more Hy's Law cases, so the right lower quadrant of the eDISH x-y plot was dubbed by me as "Temple's Corollary." It needs to be emphasized that simple elevations of ALT are not serious if not followed or accompanied by any liver dysfunctional effects such as hospitalization, inability to work, jaundice, prolonged prothrombin time, secondary kidney failure or brain obtundation. They are frequently handled by adaptation of that person's liver so it becomes tolerant of the drug, and the injury does not progress to serious dysfunction. This occurs in a great majority of patients who show some initial elevation of ALT activities when the person is first exposed to a new drug. The rate of injury worsening is a concern, because no practical monitoring program can be expected to detect liver injuries that become irreversible within one monitoring interval, such as a month as shown by a few drugs such as troglitazone, telithromycin, and a few others.

A helpful step forward might be to get rid of incorrect and misleading terminology when we discuss these issues. The term "liver function tests" is commonly applied to measurements of serum enzyme activities, such as ALT, AST, ALP, and others; this is wrong, because those tests do not measure any function of the liver whatsoever. Serum enzymes such as ALT may come from many tissues, not just the liver, and are well known to be quite unspecific, although fairly sensitive indicators of the rate and extent of hepatocellular injury. Only the TBL and INR, but not enzymes, are included in the MELD (model for end-stage liver disease) used to determine when overall liver function

is compromised sufficiently to require liver transplantation to avoid death from liver failure [26]. The scoring system, originally developed at Mayo Clinic, has stood the test of time for over 20 years, although minor fine-tuning has been suggested [27].

Another term that should be relegated to the waste basket is "Hy's Law chemistries," often seen in regulatory submissions, developed and used wrongly by some company statisticians to "diagnose" Hy's Law simply by noting peak values of $\{ALT > 2 \times ULN \text{ AND } TBL > 2 \times ULN\}$ on the step-one eDISH-like plot. The eDISH program was not designed for use as a statistical tool, but was developed for clinical reviewers who could exercise skills in medical differential diagnosis to arrive at the most likely or probable *cause* of the abnormal findings, not necessarily DILI. We are currently working on an update of the program as "eDISH2" and initiating steps to make it publicly available as a government invention.

Another example of unfortunate terminology used by the NCI Common Terminology Criteria for Adverse Events is "life-threatening" for serum enzyme elevations $> 20 \times ULN$. A vast amount of eDISH data show that such elevations of ALT are not life-threatening, are often asymptomatic, not necessarily predictive of serious hepatotoxicity, and misleading. The CTCAE to date have been based only upon expert opinions or committee votes, but not on analyses of real data. That can now be corrected by simply looking at and analyzing the vast amount of clinical trial data now in the archives of the FDA, from accumulations of many clinical trials over decades. Those data have been used for regulatory decisions, but not for the wealth of information they might yield if resources were made available to use the data for clinical research.

There is still a wide gulf of misunderstanding between many basic scientists and clinicians on the interpretation of test results. This is understandable because of the uniquely special training and experience of medical students, trainees, and practitioners in the art of medical differential diagnosis of probable causality, a skill not taught to or usually learned by chemists, pharmacologists, toxicologists, and statisticians. Those basic scientists have had great success in developing ever more effective methods to screen for drugs likely to cause hepatotoxicity in various tests systems *in vitro* and in animals. Finding potentially dangerous drugs, however, is only half the problem. The other half: drugs that appear to be reasonably safe in pre-clinical studies still may be harmful to some few patients who receive them. There are limits to how large and long controlled clinical trials can be, because of the enormous costs involved. Those necessarily limited clinical trials cannot be expected to discover the rare patients who react differently than most to new drugs, and who are susceptible to sustaining progressive liver injury and dysfunction from

the drug if it is continued too long and they are unable to develop tolerance by liver adaptation. There is no present way to identify whom those rare persons may be other than by careful, serial observation. There still is no magic biomarker to identify them. This is the medical side of the problem, and needs to be solved by clinicians who understand it, not depending on preclinical scientists to come up with easier methods.

A final suggestion for consideration by readers of this review is for a possible alternative to the conventional approval of a new drug as “safe” or disapproval as “not safe.” That simplistic idea does not take into consideration that there are many degrees of safety problems affecting the liver, from transient, asymptomatic, non-progressive increases in serum enzyme activity to the most blatant form of rapid liver failure and death or need for transplantation. Different efficacy is also seen, and not all patients show the same amount of benefit from a new drug. One dose does not necessarily fit all, despite it being easier to market if that is accepted. We all hope that the new drug will cause more benefits than harms in those receiving it, terms too often used loosely and qualitatively. A more exact quantitative evaluation requires learning of *how many* users will show *how much* benefit or harm, *how soon* and *how likely* the effects are attributable to the drug and not to natural process or other causes. Chances of benefits should exceed risks or harms if a new drug is to be useful. The quantitative use of benefits and harms should always be expressed as a **difference** and never as a *ratio*. The mis-used term benefit/risk ratio, or its inverse, do not take into consideration that some drugs have no risk (zero) in many people, or that the drug may show no benefit (zero) in some patients. Division by zero gives an infinite or indeterminate value, and should not be used.

Controlled clinical trials in large numbers of subjects, often for some years, are exceedingly costly in money spent and time off-patent consumed. Is there perhaps some alternative within the law and regulations that might permit learning more about a new drug when clinical trial data raise some suspicion of possible rare toxicity, or the drug is a member of a dangerous class of drugs? This issue deserves more thinking, debate, and work toward a solution. It is not enough just to hope-for-the-best and depend on post-marketing surveillance to discover the truth. It remains better to *prevent* such occurrences before they occur, or to devise ways to detect oncoming serious drug-induced injuries by close observation of treated patients and intervention before the injuries become irreversible. We have attempted to summarize many of the key considerations in the FDA Guidance of July 2009 [28] that all are encouraged to read very carefully, and that will be revised and updated from time to time.

4 Conclusion

Our thinking about detection, evaluation, and prevention of serious DILI has evolved greatly over the 15 years since the first conference was planned in the summer of 1998. The many thoughtful contributions of scores of participants at our annual conferences have been very much appreciated, with the expectation that they will continue.

Acknowledgments 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 John R. Senior, MD has no conflict of interest that is directly relevant to the content of this article.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

1. Editorial (no authors listed). Withdrawals of FDA-approved drugs raise questions. *Mayo Clin Health Lett.* 1998; 16(11):4.
2. Vidt DG, Speerhas RA, Reddy SS. Troglitazone: a new antihyperglycemic agent. *Cleve Clin J Med.* 1997;64(5):238–40.
3. Notice (no authors listed). Bromfenac marketed for short-term-pain relief. *Am J Health Syst Pharm.* 1997;54(19):2151–2.

4. Watkins PB, Whitcomb RW. Hepatic dysfunction associated with troglitazone. *N Engl J Med.* 1998;338(13):916–7.
5. Moses PI, Schroeder B, Alkhatib O, Ferrentino N, Suppan T, Lidosfsky SD. Severe hepato-toxicity associated with bromfenac sodium. *Am J Gastroenterol.* 1999;84(5):1393–6.
6. Wise J. Diabetes drug withdrawn (in Britain) after reports of hepatic events. *BMJ.* 1997;315(7122):1564.
7. Notice (no authors listed). Troglitazone withdrawn from market. *Am J Health Syst Pharm.* 2000; 57(9):834.
8. Friedman MA, Woodcock J, Lumpkin MM, Shuren JE, Hass AE, Thompson LJ. The safety of newly approved medicines: do recent market removals mean there is a problem? *JAMA.* 1999;281(18):1728–34.
9. Szarfman A, Machado SG, O'Neill RT. Use of screening algorithms and computer systems to efficiently signal high-than-expected combinations of drugs and events in the US FDAs spontaneous reports database. *Drug Saf.* 2002;25(6):381–92.
10. Reuben A. Hy's Law. *Hepatology.* 2004;39(2):574–8.
11. Zimmerman HJ. The spectrum of hepatotoxicity. (Kober Lecture 1968). *Perspect Biol Med.* 1968;12(1):135–61.
12. Popper H, Schaffner F. Drug-induced hepatic injury. *Ann Intern Med.* 1959;51:1230–53.
13. Davidson CS, Leevy CM, Chamberlayne EC, editors. Guidelines for Detection of Hepatotoxicity due to Drugs and Chemicals. [Fogarty Conference, 1978] NIH Publication No. 79-313. Washington, DC: US Government Printing Office; 1979. p. 109.
14. Watkins PB, Desai M, Berkowitz SD, Peters G, Horsmans Y, Larrey D, Maddrey W. Evaluation of drug-induced serious hepatotoxicity (eDISH): application of this data organization approach to phase III clinical trials of rivaroxaban after total hip or knee replacement surgery. *Drug Saf.* 2011;34(3):243–52.
15. Senior JR. Alanine aminotransferase: a clinical and regulatory tool for detecting liver injury—past, present, and future. *Clin Pharmacol Ther.* 2012;2(3):332–9.
16. Ledley RS, Lusted RB. Reasoning foundations of medical diagnosis: symbolic logic, probability and value theory aid our understanding of how physicians reason. *Science.* 1959;130(3366):9–21.
17. Vecchio TJ. Predictive value of a single diagnostic test in unselected populations. *N Engl J Med.* 1966;274(21):1171–3.
18. Qureshi ZP, Seoane-Vasquez E, Rodriguez-Monguio R, Stevenson KB, Szeinbach SL. Market withdrawal of new molecular entities approved in the United States from 1980 to 2009. *Pharmacoeconom Drug Saf.* 2011;20(7):772–7.
19. Chen HJL, Bloch KJ, Maclean JA. Acute eosinophilic hepatitis from trovafloxacin. [Letter]. *N Engl J Med.* 2000;342(5):359–60.
20. Lucena MI, Andrade RJ, Rodrigo L, Salmeron J, Alvarez A, Lopez-Garrido MJ, Camargo R, Alcantara R. Trovafloxacin-induced acute hepatitis. [Letter]. *Clin Infect Dis.* 2000;30(2):400–1.
21. Rockey DC, Seeff LB, Rochon J, Freston J, Chalsani N, Bonacini M, Fontana RJ, Hayashi PH. Causality assessment in drug-induced liver injury using a structured expert opinion process: comparison to the Roussel-Uclaf causality assessment method. *Hepatology.* 2010;51(6):2117–26.
22. Danan G, Benichou C. Causality assessment of adverse reactions to drugs—I. A novel method based on the conclusions of international consensus meetings: application to drug-induced liver injuries. *J Clin Epidemiol.* 1993;46(11):1323–30.
23. Benichou C, Danan G, Flahault A. Causality assessment of adverse reactions to drugs—II. An original model for validation of drug causality assessment methods: case reports with positive rechallenge. *J Clin Epidemiol.* 1993;46(11):1331–6.
24. Teschke R, Frenzel C, Wolff A, Eickhoff A, Schulze J. Drug induced liver injury: accuracy of diagnosis in published reports. *Ann Hepatol.* 2014;13(2):248–55.
25. Kleiner D, Chalasani NP, Lee WM, Fontana RJ, Bonkovsky HL, Watkins PB, Hayashi PH, Davern TJ, Navarro V, Reddy R, Talwalkler TA, Stolz A, Gu J, Barnhart H, Hoofnagle JH. Drug-induced liver injury network (DILIN). *Hepatology.* 2014;59(2):661–70.
26. Dickson ER, Grambsch PM, Fleming TR, Fisher LD, Langworthy A. Prognosis in primary biliary cirrhosis: model for decision making. *Hepatology.* 1989;10(1):1–7.
27. Kim HJ, Lee HW. Important predictor of mortality in patients with end-stage liver disease. *Clin Mol Hepatol.* 2013;19(2):105–15.
28. Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, Food and Drug Administration. Guidance for industry. Drug-induced liver injury: premarketing clinical evaluation. *Drug Saf.* 2009. <http://www.fda.gov/downloads/drugs/guidancescompliance/regulatoryinformation/guidances/ucm174090.pdf>.