Section 2
Hepatic Side Effects
5
Drug-Induced Liver Injury: Clinical and Diagnostic Aspects

John R. Senior

5.1
Introduction

This chapter is divided into four main sections. This section deals with introductory and general comments, and isoniazid hepatitis as an example, Section 5.2 provides a discussion of postmarketing concerns, Section 5.3 deals with new development issues, and Section 5.4 provides closing comments and research opportunities.

It is pertinent first to ask the following question:

Is it a toxic drug? Or an especially susceptible patient?

We must remember and accept that drugs have different effects on different people, both for intended beneficial effects on a targeted organ or receptor and for possibly adverse effects on unintended, off-target sites. These adverse effects, here called “antitarget” effects, may occur. The latter term is relatively new, appearing first in a published title only as recently as 2005 [1]. It refers to a receptor, enzyme, or other biological site for a drug action that leads to undesirable or adverse side effects when the site is affected. Perhaps we should consider extending that definition to an especially susceptible person.

The problems facing those who develop new drugs and those who later prescribe them are quite different. Drugs are developed and selected to induce beneficial and not harmful effects in patients to whom the drug is meant to be prescribed, recommended, or sold. It is not enough just to develop new drugs that are safe for most or nearly all of the people, because the very same dose of

*Although the material discussed here is based largely on experiences since 1995 as a medical reviewer and since 2000 as a consultant in hepatology at the FDA, the opinions expressed do not necessarily reflect regulatory positions or policies of the agency. The author received no special funding for preparing this chapter and has no conflicts of interest.
the very same drug may be quite safe, well tolerated, and effective for most patients, but quite toxic to a few. Not just the number of effects, but their severity and consequences, time course, and degree of causal likelihood should be considered. Just a few patients who show serious adverse effects, even if rare, may negate moderate benefits to many. This balance may determine whether or not a drug remains available in the market. It is also easy to appreciate why liver injury is so often seen after exposure to a drug, for the liver receives something over a quarter of the cardiac output of blood, including both portal venous and hepatic arterial flow, and metabolizes xenobiotic nonendogenous compounds as well as many endogenous materials. Therefore, liver cells are exposed to the fresh metabolites that may cause injury or perturbation by a variety of mechanisms that are still being explored by basic scientists, including many authors of other chapters in this book. The liver is of course not the only antitarget for adverse drug effects, but has been a principal one [2], perhaps because of its metabolism of drugs and compounds. Severe hepatotoxicity was the leading cause for withdrawing approved drugs from the market previously, before cardiovascular problems supplanted it.

The liver is unique among our major internal organs in having the remarkable ability to regenerate even after quite drastic resection or damage to two-thirds or even more of its volume, regrowing quickly to its original size and function, and then stopping, as so well summarized in 1967 by Dr. Nancy Bucher in her excellent two-part review [3,4]. A very recent update on liver regeneration was written by Itoh and Miyajima [5]. The unique ability of the liver, among all our vital internal organs, to regrow rapidly after injury has been known for a very long time, as recalled to us by the story of Prometheus [6]. The ancient Greeks were onto something, in naming the liver “hēpar” (after hēpaomai, which means “to repair oneself”); they clearly understood the liver’s capability to regenerate.

Short of full regeneration, hepatic cells often appear to be able to adapt and change their functioning, becoming tolerant to drugs initially injurious. They may show hepatocellular necrosis or apoptosis first, and then little or no adverse effect when re-exposed to drug [7].

5.1.1
Postmarketing Hepatotoxicity versus Hepatotoxicity in Development

Although the focus of this book is on identifying and excluding from development new compounds that are likely to be toxic to the liver, a persisting major problem is the often rare but serious hepatotoxicity that occurs after marketing in certain individual patients, a phenomenon termed “idiosyncrasy,” for lack of better understanding of its mechanisms or causes. The often-used distinction between “dose-related” and “idiosyncratic” drug reactions is not absolute. There are very few drugs at the extreme ends of the Zimmerman [8] “spectrum of hepatotoxicity,” shown as Figure 1.1 in his texts of 1978 and 1999 [9].

Most adverse drug reactions manifest features of both (see Figure 5.1). It is overly simplistic to state that some drugs are predictably and reproducibly “true”
hepatotoxins because of properties that cause dose-related liver injury, whereas others show none of those features and effects depend only upon undefined idiosyncrasies of the person to whom the drug is given. Although there are many forms of chemically induced liver injury, idiosyncrasy of the host is responsible for many of the cases of serious hepatotoxicity currently seen in new drugs after approval for marketing. The term means “one’s own mixing together” (idios = one’s self + syn = together + krasis = mixture, a word of Greek derivation) and implies that each human is somewhat different from others despite many common characteristics.

5.1.2
Isoniazid – If It Were Newly Discovered, Would It Be Approved Today?

Hydrazine derivatives of pyridine carboxylic acids have been known for more than 100 years, dating to 1912 when two graduate chemistry students in Prague [10] synthesized isonicotinyl hydrazide, known as isoniazid and abbreviated INH. Isonicotinic acid is the 4-isomer of nicotinic acid (vitamin B₃, 3-carboxypyridine). Free hydrazine is a toxic, flammable, and unstable substance used for rocket fuel and for emergency power in F-16 fighter jets. Nicotinic acid, vitamin B₃, was called niacin (= nicotinic acid vitamin) to distinguish it from the addicting plant alkaloid nicotine, found by Jean Nicot in tobacco, but also in other
plants such as green pepper, eggplant, and tomato (see Figure 5.2). It is also toxic (lethal dose for 50% of humans is 0.5 mg/kg, 100 times the toxicity in rats).

In 1936, Gerhard Domagk in Germany, after researching many compounds, discovered sulfanilamide as the first antibacterial agent. He was awarded the Nobel Prize in 1939, but the Nazi regime would not allow him to accept it. He later did so in 1947 but without the monetary award. In 1944, Schatz and Waksman discovered streptomycin to be active against tuberculosis [11], a first pharmaceutical alternative to the regimen of sanatorium rest and fresh air, leading to another Nobel Prize, for Waksman in 1952. Domagk kept on working and searched among many compounds for activity against tuberculosis. In 1950, he found active thiosemicarbazones [12] and isoniazid, an orally administrable compound found to be more active against tuberculosis by Grunberg and Schnitzer [13].

INH was first used for treatment of tuberculosis in June 1951 by Selikoff and Robitzek [14] at Sea View Hospital on Staten Island, NY, as cited by Lapp et al. [15], who wrote that “A new drug is with us in the treatment of tuberculosis. It is a synthetic chemical of known composition, relatively non-toxic, cheap, and simple to prepare.” INH use spread rapidly in 1952 in North America and Europe [16], because it could be given orally.

An early case of serious liver injury with jaundice was published in 1953 [17]. The later story of clinical liver effects of INH has been well summarized at the LiverTox website [18], with 363 references. INH became used mainly to prevent tuberculosis. For treating active disease it had been found that INH alone was insufficient, and combinations of two or more drugs (such as pyrazinamide, para-aminosalicylic acid, rifampin, ethambutol, and streptomycin) were needed to prevent emergence of resistance. Sporadic reports of a few deaths in patients taking just isoniazid appeared in the 20 years after it was first used to treat
tuberculosis, sometimes complicated by underlying disease, administration of other antitubercular therapy, but in some cases isoniazid alone appeared capable of causing liver injury and sometimes killing susceptible patients [19–22].

Two notable studies were carried out in the early 1970s. First was a Public Health study of 13,838 patients at 26 United States health clinics from July 1971 to November 1972 [23,24]. Black et al. reported [23] that 224 of the 13,838 patients (1.6%) were suspected to have INH-caused hepatitis on clinical grounds, but no routine testing of serum enzymes was done consistently. Of that number they assessed 87 as probably INH-induced hepatitis, another 76 possibly related, 18 due to other liver disease, 20 as not hepatitis, and 23 with insufficient data to decide. They then added another 27 cases from other hospitals (with no known total exposed) to make up 114 cases probably induced by INH, to characterize the syndrome. There were 7 deaths from the 87 Public Health group and 6 from the 27 added patients, totaling 13 fatal cases from the 114 probably induced by INH (1.1%). The clinical picture of probably INH-induced hepatitis was chemically and histologically not distinguishable from viral hepatitis but was more likely to occur among elderly patients. Fatal outcome was more likely if INH therapy was prolonged for more than 2 months.

Later analysis of the large Public Health study by Kopanoff et al. [24] reported 8 deaths, 7 in Baltimore of 3,196 patients (0.2%), among 25 cases of INH-induced hepatitis (0.8%). Six of the seven Baltimore cases were female, five black and one white, and the seventh a white male. One death occurred in Cincinnati, out of 709. No deaths and 64 cases (0.6%) of probably INH-induced hepatitis occurred among 9,933 patients from the other 24 cities.

The second study [25] was smaller and more controlled, done at the Eastern State Hospital (a psychiatric hospital) in Williamsburg, VA, of 358 patients in 1971–1972. Of them, 218 were reactive to tuberculin and were treated with isoniazid 300 mg/day as prophylaxis against tuberculosis and 140 nonreactive patients were followed but not treated with INH. In that more carefully monitored study of the psychiatric patients, 201 of the 218 patients given INH completed at least 4 weeks of a planned year of observation with regular monthly testing of serum aspartate aminotransferase (AST) and total bilirubin (TBL), as did 107 of the 140 untreated patients. Frozen sera were sent to the Center for Disease Control in Atlanta, but neither were analyzed immediately nor influenced clinical management. Weekly records were kept of symptoms and medications. Chemistry results were not known until after the treatment year. In 22 of the 173 patients (12.7%) who completed a year, AST elevations were \( >30 \text{ mU/ml} \), taken as the upper limit of normal (ULN). Three (see Figure 5.3) showed elevated TBL \( >3 \text{ mg/dl} \) in addition to AST elevations from 14.6 to 30.7 times the ULN – what we would now consider Hy’s law cases if caused by INH and not cholestatic!

None of the patients receiving INH had any clinical symptoms during treatment and all three with highest abnormalities recovered to normal levels of bilirubin by the end of their treatment despite continuing the INH, and did not progress to liver failure. Note that the older patient (M61b) showed delayed rise
in AST and persisting but lesser elevations, although his bilirubin had returned to normal when the AST subsided.

These findings then confirmed what had been shown beautifully at New York Hospital in 1969 by Scharer and Smith [20] – that INH alone could cause acute liver injury, with fairly frequent elevations in serum transaminases, but less common liver dysfunction with bilirubin elevations. Patients showing worrisome abnormalities did not usually progress to liver failure or death. In very large studies, with many thousands of patients observed, some deaths could occur, particularly if patients were not closely followed. These were important lessons for a valuable drug that prevented tuberculosis!

The issues were argued back and forth, but a large study by Nolan et al. [26] of 11 141 patients treated with INH to prevent tuberculosis in Seattle–King County from January 1989 through December 1995 did much to resolve the debate. They also observed another 1427 patients with active tuberculosis treated with combination regimens. Patients were seen in clinic monthly, given medication refills, and were reminded at each visit to take note of and report symptoms of liver toxicity, but were not monitored with serum tests. Only 11 of the 11 141 patients on INH prophylaxis (0.10%) had symptoms leading to findings of elevated serum hepatocellular enzyme activities (peak AST: 303–1160 U/l). Among them, eight also showed elevated bilirubin (2.8–10.6 mg/dl), and one was hospitalized. None developed liver failure or died, and all recovered without sequelae. Among the 1427 patients treated with multiple antitubercular drugs for active disease, there were 15 (1.05%) who showed hepatotoxic effects. They attributed these results to including mostly (80%) only younger patients <35 years of age and using clinical symptoms to trigger investigation rather than serum enzyme tests.
Mitchell et al. had postulated [25] that people who showed rapid acetylation rates might be at higher risk of developing hepatitis from isoniazid, whose principal metabolite was acetyl isoniazid [27], competitively inhibited by other substrates [28], and that acetylation was polymorphic [29]. However, this was challenged by several other investigators, leading to the current consensus that slow acetylators may be at higher risk [30–32].

As may be seen, it is necessary to consider not only the first step of INH acetylation, but also its subsequent hydrolysis that can generate even more toxic secondary metabolites, such as acetylhydrazine and even free hydrazine (see Figure 5.4). Rapid acetylation to harmless, nontoxic diacetylhydrazine is protective, so slow acetylators may be at greater risk.

This example illustrates what could have been a major problem, if it were judged by current standards and criteria. There were many “Hy’s law” cases that today might be ominous for a new drug, but over that earlier time it became evident that isoniazid was doing far more good than harm, and that it could be used safely. We can only speculate on what the answer might be to the question posed at the beginning of this section.

5.2 Special Problems of Postmarketing Hepatotoxicity

There are major differences between pre- and postmarketing strategies for dealing with hepatotoxicity. Clinical studies may include hundreds or even thousands of subjects, but after marketing, hundreds of thousands or millions of patients may be given or take the drug, enormously increasing the likelihood of seeing a relatively rare but serious adverse event. In addition, patients to whom the drug
is prescribed after approval or perhaps taken without prescription may be less carefully selected than subjects who had met protocol restrictions of controlled studies before marketing, and may be less carefully followed. Adverse drug reactions are often missed or never reported if they occur, or if reported too frequently are not well enough described for making accurate diagnosis of the probable cause. Errors of both omission (false-negative, beta errors) and nonspecificity (false-positive, alpha errors) are common in postmarketing reports.

During clinical development, in controlled trials under protocols, subjects chosen for participation are usually observed for liver cell injury by scheduled monitoring of serum enzyme activities that are quite sensitive for detecting early injury. In the postmarketing use of approved drugs, the focus is on efficacy, whether the drug administered produces beneficial effects, and monitoring is usually poorly done or not at all. Even if elevated serum enzymes are found, they may not be well investigated to determine the severity and cause of the findings, or the results not reported. Consequently, adverse, off-target effects are too often missed, not detected, and underreported, creating a large type 2 (beta, false-negative) error, and no reliable assessment of the incidence of the problem is possible. In clinical trials, where routine monitoring is done, there is more likely a type 1 (alpha) error in which modest elevations of serum enzyme activities are taken as important but represent false-positive findings that would resolve without treatment or interruption of treatment, due to the frequent process of hepatic adaptation.

5.2.1 Voluntary Monitoring after Approval for Marketing

In some cases, findings of modest liver injury during controlled clinical trials may lead to labeling recommendations for periodic monitoring of serum amino-transferase activity to detect early hepatocellular injury. This has often turned out to be ineffective because it is dependent on cooperation of patients and their physicians, both of whom grow weary of receiving repeatedly negative test results, because what is being looked for is rare, and may not occur for some time, so that the vast majority of test results are negative. This leads to quitting the monitoring, which consequently cannot be effective. Alternatively, waiting for patients to report symptoms such as nausea, fatigue, or jaundice may cause such delay that really serious liver toxicity by then may be irreversible even if the drug is stopped. No universally satisfactory alternative to routine monitoring has been found as yet. Theoretically, patients under treatment should be closely observed, and even early and mild symptoms investigated thoroughly, but this is extremely difficult to enforce. This situation has been a driving factor for seeking predictive biomarkers.

5.2.2 Prediction of Serious, Dysfunctional Liver Injury

Unfortunately, there are not as yet any biomarkers that reliably predict which individual will develop progressively severe liver injury from a drug. This is
especially difficult because of the rarity of serious effects leading to disability, life-threatening illness, death, or need for transplantation. Only observation over time currently is effective in finding an individual who can neither tolerate nor adapt to the drug, and the only treatment is to stop administration of the drug to those persons. Because of the quite common ability of many patients to adapt to a new drug after showing transient or nonprogressive liver injury from it, negative rechallenge results are not convincing proof of susceptibility to serious drug-induced liver injury (DILI). Positive rechallenge, on the other hand, is powerful evidence of such susceptibility, but is dangerous unless patients are very closely watched. Not all prescribing physicians in medical practice are experts in liver disorders, and their ability to recognize, investigate, and diagnose the problem is quite variable. Lack of reporting, or inadequate reporting of necessary details, is too common, and no reliable estimate of the incidence of drug-induced liver injury can be made from the spontaneous and often isolated, delayed reports that reach publication or official recognition.

5.2.3 Severity of Liver Injury Is Not Measured by Aminotransferase Elevations

It is important to appreciate that severity of drug-induced injury to the liver varies from mild asymptomatic injury detectable only by elevated serum aminotransferase or other biomarkers without symptoms or functional disturbance of the liver to acute liver failure and death or need for transplantation. Activities of serum enzymes, released from injured hepatocytes, are not valid measures of the severity of adverse liver effect, which is really dependent on how much of the whole organ is not injured and still able to function well. The level of serum activity of enzymes released from injured cells depends upon the rapidity and extent of the injury and is a useful indicator of the urgency to investigate the patient to determine the course and cause of the abnormal findings. The only tests commonly and routinely done that truly measure a liver function are the serum bilirubin concentration (measures ability of liver to clear plasma of bilirubin, conjugate it with glucuronide, and excrete it into bile) and the prothrombin time or its derived international normalized ratio (INR; measures ability of the liver to synthesize and regulate the level of a key coagulation protein). Serum albumin concentration is too insensitive to be useful for acute cases. Other tests of liver functions may be done in special cases, but regulation of serum aminotransferase activity is not a liver function, so activities of biomarker enzymes in serum should not be called “liver function tests.”

5.2.4 Attempts to Standardize Terminology

International conferences were held in 1974 and 1978 at the Fogarty International Center of the National Institutes of Health (NIH) in Bethesda, MD, on standardization of liver and biliary disease nomenclature [33] and detection of
drug-induced hepatotoxicity [34]. At the latter conference, it was the consensus (not data-based, but opinions of experts) that elevations of serum enzyme activities were “markedly abnormal” if more than three times the upper limit of the normal range (>3× ULN). This concept was carried forward to the first 1982 version 1 of the National Cancer Institute Common Toxicity Criteria at which a slightly lower cutoff of >2.5× ULN was taken as distinguishing mildly from moderately severe toxicity (grade 1 from grade 2). The Common Toxicity Criteria were revised in 1998 (version 2) and 2006 (version 3), at which time the listing was greatly expanded and the name changed to Common Terminology Criteria for Adverse Events (CTCAE) and the cutoff between grades 1 and 2 was raised slightly to 3× ULN.

The CTCAE are again in revision for version 5, expected in 2015, but continue to use the concepts of earlier versions. Standard liver tests, such as serum enzyme activities of alanine and aspartate aminotransferases (ALT and AST) and alkaline phosphatase (ALP), as well as concentration of TBL, are graded by how much they are elevated above the ULNs. Grades of severity are assigned, depending on how much elevated: grade 1, mild and not requiring intervention, from >1 to 3× ULN for ALT and AST, from >1 to 2.5× ULN for ALP, and from >1 to 1.5× ULN for bilirubin. Grade 2 denotes moderate severity, for up to 5× ULN for the three enzymes and up to 3× ULN for bilirubin. Grade 3 elevations are taken as severe, with need for or extension of hospitalization, disability but not immediately threatening to life, for enzyme elevations from >5 to 20× ULN or bilirubin elevations from >3 to 10× ULN. Grade 4 enzyme elevations of >20× ULN or bilirubin elevations of >10× ULN are classified as immediately life-threatening and requiring urgent intervention. Grade 5 is used for fatal outcomes, not applicable to these serum tests. These grades were not based on data analyses but on opinions from unidentified committees. Unfortunately, no definition of normal was made, and the range of normal was left to local laboratories to define, which has led to wide variation in the putative normal ranges. Even greater than differences in methods used to measure enzyme activity are differences in people who are assumed to be normal because they do not appear ill but may have undiagnosed diseases.

5.2.5
What Is the “Normal” Range, or the “Upper Limit of Normal”?

The word normal has many meanings. Two that are most pertinent to this discussion are the statistical term meaning frequent or common, and the medical term meaning well or healthy, not diseased. After marketing, major interest is whether the findings in a given patient should raise concern about a possible disease being present. In clinical practice, the medical view should be used, certainly to assess individual cases. An approach to the question was taken by

Dr. Daniele Prati [35] who studied prospective first-time blood donors in Milan, ruling out those who had asymptomatic chronic hepatitis C, excessive use of alcohol, obesity, skeletal or cardiac muscle damage, concurrent medications, or other possibilities, concluding from 3927 volunteers that healthy normal upper limit of serum ALT normal for men was 30 U/l ($\mu$mol/min) and for women was 19 U/l. She also reported median values of the tests for healthy men and women as 13 and 9 U/l, respectively. These are much lower ULNs than found by statistical methods that ignore elevations caused by disease, age, and gender. There are no known clinical implications for having lowered serum enzyme activities.

5.2.6 Diagnostic Test Evaluation

Evaluation of diagnostic tests for distinguishing diseased from nondiseased persons was advanced greatly in 1966 by the much-cited paper of Thomas Vecchio [36]. It introduced the idea of “positive predictive value” and “negative predictive value” of the test result. It is unfortunate that Vecchio used the word “predictive” for a single test result, without knowing what the results were before or might be afterward. However, he emphasized a critical point that has been overlooked too often, which was the prevalence of the sought disease in the population considered and tested (see Figure 5.5). For occurrence of a new finding, onset of drug-induced liver injury at a future time, the term incidence is better.

It is evident that evaluation of such a test looks very good indeed, when results from equal numbers of people with or without the disease are compared. However, when the results of such a very well “validated” test are applied to seeking a relatively rare event in a population where only 1 per 1000 develops disease (compare results above), a very different conclusion would be reached about the value of a positive test result (Figure 5.6).

Obviously, a positive test that is nearly always wrong is not useful, and the value of test result must consider incidence of new events or prevalence of existing problems.

<table>
<thead>
<tr>
<th>test result</th>
<th>SDILI present</th>
<th>none</th>
<th>total count</th>
<th>“value”</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>47,500 TP</td>
<td>2,500 FP</td>
<td>50,000 P</td>
<td>0.95 looks good</td>
</tr>
<tr>
<td>negative</td>
<td>2,500 FN</td>
<td>47,500 TN</td>
<td>50,000 N</td>
<td>0.95 looks good</td>
</tr>
<tr>
<td>incidence</td>
<td>50,000 50%</td>
<td>50,000 50%</td>
<td>100,000 tested</td>
<td></td>
</tr>
<tr>
<td>sensitivity</td>
<td>sensitivity, 95%</td>
<td>specificity, 95%</td>
<td>accuracy, 95%</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.5 Validation for common (50% incidence) events.
5.2.7 Determination of the Likely Cause of Liver Abnormalities

The most notable deficiency in postmarketing reports of liver injuries possibly caused by a drug or chemical is lack of proof of causality. We have come to realize that almost any drug may cause at least some transient injury to the hepatocytes of at least a few people, and that the range of the injury is from minimal, asymptomatic, slow leakage of enzymes from the injured cells into circulating plasma all the way to liver failure. Although those leaked enzymes do not have any useful function in the plasma and are simply proteins awaiting proteolytic degradation, their enzymatic activity can be coaxed out by addition of cofactors and substrates for some days before they are degraded and inactivated. We also have come to appreciate that not only hepatocytes but other types of liver cells may also be injured, and that the patterns of injury are diverse, mimicking almost all types of liver diseases. Consequently, it is not possible to diagnose injuries as drug-induced, actually caused by the drug, on liver biopsy. At best it may be said that patterns of histological findings are “compatible with” or “possibly related to” the drug effect. Proof of causality is not yet possible by any single biomarker. Diagnosis of drug-cause requires obtaining extra supplemental clinical information to rule out to a reasonable degree of likelihood other possible causes. At present, this cannot be done by serum chemistries alone, but requires medical differential diagnosis. It remains unfortunate that drug-induced liver injury is still a “diagnosis of exclusion” that requires medical expertise. It seems a waste of effort to pursue minor forms of liver injury that produce no symptoms or loss of true functions of the liver, to chase after transient serum enzyme increases. That effort should be put on observation and investigation of potentially serious cases of DILI that lead to clinical illness, inability to work, hospitalization, or worse.

At present, diagnosis of the cause of liver injury is still a medical art and cannot be done accurately by biomarkers or nonphysicians. It is not sufficient to use the weaker and almost meaningless terms of “associated with” or “related to” to a
5.2.8

Treatment and Management of DILI in Practice

Treatment is dependent on correct diagnosis; wrong diagnosis leads to wrong treatment. If the injury is mild and transient, no treatment may be required, and administration of the drug may be continued, if it is having a beneficial effect. If the injury is worsening, it is wise to interrupt treatment until the course over time is clearer, and to be cautious in restarting the drug. If the injury becomes serious, the drug must be stopped. There are no known antidotes that are generally useful, so the difficult burden of deciding whether or not to stop drug administration falls upon the prescribing physician. Serious DILI may come rapidly, and is unpredictable. Despite this, detection and recognition of DILI is often missed and unreported even if found. Mild DILI is often asymptomatic, detectable only by testing serum for high activity of enzymes such as ALT, AST, and ALP.

5.3

Special Problems for New Drug Development

As mentioned earlier, the problem of avoiding serious DILI from potential new drugs is a major concern of pharmaceutical companies engaged in the very competitive and costly process of developing new drugs. Testing for safety in human subjects taking part in controlled clinical trials requires large numbers of study sites and subjects. This makes it possible to start applying the test questions mentioned in Section 5.1: “How many? How much? How soon? How likely?” It is very difficult to find out exactly how many patients have been prescribed a drug after approval for marketing, how many actually took what dose for how long, how many had what side effects or when, or even if they did whether it was reported. Controlled clinical trials provide an opportunity to do so, because all subjects need to be accounted for, routine observation with mandatory reporting is done, and records are kept. Opportunity is available to obtain some idea of the incidence of adverse effects and to determine whether they are mild, moderate, or severe; when they occur after administering the drug; and to investigate the serious cases for probable cause (which may not be the drug). We shall consider each of these four simple questions in following sections.

5.3.1

How Many?

The incidence of adverse effects is a useful and important issue that is difficult or even impossible to determine with any reliability after marketing. Only during
controlled trials when subjects are carefully selected, followed, data recorded, and analyzed can some idea of how often abnormalities occur (incidence), at which doses of medications in which populations, that is, the incidence of the problem, be obtained. The need to obtain a quick bird’s-eye view of all the subjects in one study was a force for developing our eDISH program [37], acronym for evaluation of drug-induced serious hepatotoxicity. The term “serious” refers to cases with liver dysfunction sufficient to cause clinical disability, to require or prolong hospitalization, or to cause death or need for transplantation. More will be said about eDISH later. It should be noted that eDISH provides only raw estimates of how many subjects had how much elevation of peak serum ALT activity and total bilirubin concentration in the initial $x$–$y$ log–log plot of ALT values on the abscissa and TBL on the ordinate. Further analyses for time course of individual subjects and probable causality by supplemental clinical narrative information must be done to reach conclusions.

5.3.2

How Much?

The severity of drug-induced liver injury may be suspected but not proved by the fairly sensitive but not specific serial measurement of serum ALT activity. Because of its high sensitivity, few cases are missed, so the false-negative (beta, type 2) error is very low. But not all elevations of ALT are due to drug-induced injury, and must be distinguished from other possible causes such as viral infections, skeletal or cardiac muscle injury, and other liver diseases. Without considering the other possible causes, which requires differential diagnosis, grossly high false-positive (alpha, type 1) error rates occur. This is addressed by the inclusion of serum TBL concentration, a true liver function test, because only the liver clears bilirubin from the plasma, conjugates it with glucuronide, and excretes it into the bile. Most of the bilirubin originates from red blood cell turnover that releases the iron-containing, tetrapyrrolic red pigment heme [38] that is then oxidized by ring opening, salvage of the Fe$^{2+}$, and production of biliverdin, and then reduced to bilirubin and transported bound to serum albumin [39] to the liver sinusoids where it enters hepatocytes and is carried by intracellular proteins for conjugation [40].

It is not a liver function to regulate the activities of serum enzymes that themselves have no function in plasma, but only had some function inside the cells from which they were released because of injury to those cells. The clinical severity of liver disorders depends upon how well the uninjured hepatocytes can function after loss of some by injury. It is wrong to conclude that extent of injury determines remaining function rather than still functioning uninjured cells. True functions unique and specific to the liver include the clearance of bilirubin from the plasma and synthesis and regulation of proteins such as the coagulation factors (namely prothrombin) and albumin. The plasma albumin, the principal protein of plasma, is in the range of 100–150 g, and changes rather slowly, so it is not a sensitive indicator of acute liver injury. It is wrong to assess severity of liver dysfunction by grading elevations of serum ALT values.
5.3.3  
**How Soon?**

The time between first administration of a drug and the hepatic response to it is extremely variable, ranging from days to months. It probably depends on just what mechanism may be causing the injury, an interesting field of research that is still underway, depending on what metabolites are produced and the host responses to them. There is a vast amount of basic science study underway that will be important to understanding the latency between drug administration and antitarget effect observable. Some of this is covered in the other chapters of this book, and more will come.

5.3.4  
**How Likely?**

This is one of the most difficult issues facing developers of new drugs. It goes back to the opening question of whether an observed antitarget or adverse effect such as liver injury was caused by the drug or by some peculiarity of the person receiving it. As stated above, it is not enough to say “associated with” because the effect was seen and the drug was given. The somewhat stronger term of “related to” implies some cause is suspected, but not convincingly. To state that the effect was drug-induced means that other causes have been excluded, which requires work to consider what all of those causes might be, and what needs to be done to exclude them, for we do not yet have any test or procedure to make a positive diagnosis of drug-induced liver injury. That diagnosis remains one of exclusion. Routine study protocols do not usually specify all possible causes of adverse effects, nor what must be done to make or exclude their likelihood as an explanation for the findings. This is a process of medical differential diagnosis, learned and practiced by physicians as students, and then over many years of experience improved if not perfected. It is a skill and art unique to medical scientists and not usually included in the training and experience of basic scientists in the fields of chemistry, toxicology, pharmacology, or statistics that play such important roles in new drug development. It has been a field of consultation activity for 15 years at the FDA for this writer, from which much has been learned and continues to be. It is perhaps better to address the questions before approval of new drugs for use in medical practice, because more detailed and better information is available, rather than try to collect information from busy practitioners after approvals.

5.3.5  
**Compared with What?**

A fundamental issue is whether a given subject under observation should be compared with a group of normal people, however that may be determined, or to him/herself, in looking for drug effects. This has come into prominent
consideration for evaluating many new treatments proposed for chronic hepatitis C, where many or most of the patients treated and studied will show abnormalities before treatment starts, and then often show a reduction in disease activity when the viral infection is suppressed. This effect may lead to a new and lower “baseline” level of serum enzyme activity with which to compare later effects during treatment. If serum ALT rises without viral load increase, it may indicate possible drug toxicity. It may be preferable to establish the individual patient’s own baseline for liver injury and dysfunction biomarkers, rather than comparing them with a hypothetical normal population. This issue has been discussed very recently by Dr. Leonard Seeff, an established expert in viral hepatitis C [37], who pointed out that it is time to rethink the current acceptance of “normal” ranges provided by machine vendors or local laboratories and consider using changes in each person’s own “baseline” to evaluate drug effects.

5.3.6 ROC Curves

It has become very popular to evaluate tests with receiver operating characteristic (ROC) curves, based on comparing test sensitivity (true-positive results) on the ordinate with the complement of specificity (false-positive results) on the abscissa. The origin of this method dates back to World War II when it was very important to evaluate signals of radar reflection to distinguish approaching enemy planes from flocks of geese or other harmless results that would lead to the unnecessary “scrambling” of defensive aircraft, wasting precious fuel. It was later applied to psychology and radiology, and now to many areas of medicine. An example of such a curve is shown in Figure 5.7, in which the point at the maximal distance from the diagonal line of no discrimination is at (0.05, 0.95), indicating (1 – specificity) of 5% (or specificity of 95%) and sensitivity of 95%. The area under the ROC curve is 0.982, a very high value (Figure 5.7). The best point for a perfect test would be at (0.00, 1.00) and has an area under it of 1.000, indicating that all diseased persons were detected and no undiseased people were wrongly classified. Such perfection has not been achieved with available tests, and there is a well-known trade-off between seeking higher sensitivity at the cost of lower specificity, so the cutoff value separating positive from negative test results must be chosen to optimize it.

Unfortunately, prevalence and incidence are not considered in ROC curves, and therefore they should not be applied to detection of rare events if “validated” or “qualified” on common ones. Even if the same 95% sensitivity and 95% specificity are “validated” by comparing equal numbers of patients with the disease sought (here serious drug-induced liver injury) and those without disease, but applied to screening for rare events, mischief is done. ROC curves are the same for both common and rare events; they should not be used to validate new tests if neither incidence nor prevalence is considered.
5.3.7 eDISH: Especially for Controlled Trials

We have developed an in-house program for evaluating serum liver tests during clinical trials, plotting peak values of serum ALT activity and TBL concentration for all subjects in a given study during their time of observation. This program has been used at the FDA by reviewers and consultants since 2004 when it was first applied to a study of 3922 patients randomized in about equal numbers to either an experimental drug (X) or a control agent (C). The symbols are plotted as peak observed log_{10} x–y values for each subject.

These graphs were developed as an aid to medical reviewers, to help them visualize very quickly the elevations of serum ALT (abscissa) and TBL (ordinate). The data were obtained from new drug applications and plotted according to a software program written in SAS/IntrNet code by our research statistician Dr. Ted Guo (Figure 5.8). Arbitrary cutoff levels of 3× ULN for ALT and 2× ULN for TBL were chosen, in keeping with 1978 Fogarty conference [34], as being “markedly abnormal,” and with concepts of Dr. Hyman Zimmerman concerning ALT as an indicator of acute hepatocellular injury to the liver and bilirubin concentration as a measure of reduction in overall liver function.

The program was intended for internal FDA use, was not published openly, but was very quickly copied by many pharmaceutical companies who wished to evaluate their own data before submitting them for review and possible approval. The second step of eDISH programming is to point to an individual symbol on the first x–y plot to obtain all the liver test data for that single subject for the

---

entire period of observation, including ALT, AST, ALP, and TBL values, all plotted in terms of ULN multiples. This time course made it quite simple to see if the hepatocellular injury preceded the rise in bilirubin or not. If the latter, some other cause for the bilirubin retention had to be sought. The third step in the eDISH program was to request a clinical narrative report describing alternative possible explanations for the abnormal findings, to aid in the medical differential diagnosis and an estimate of how likely the leading possible causes might be.

This allowed a clinical idea of what might have been the probable cause, more likely than all others combined (for an estimate of 51–75%), a very likely cause (>75–95%), or an almost certain cause (>95%). These numbers were not exact, and were based on the judgment of the medical analyst. The value of the eDISH program to academic consultants to the pharmaceutical industry has been cited by Watkins et al. [41,42].

A frequent misuse of the eDISH program and concepts followed when statisticians of some companies began to refer to “chemical Hy’s law cases” for subjects whose peak values of ALT and TBL placed them in the right upper quadrant of the first graph. This is wrong, for the diagnosis of a Hy’s law case absolutely requires that it be “drug-induced, hepatocellular, and with visible jaundice.” It demands medical diagnosis of causality, and ruling out all other causes than the drug, and not principally cholestatic in presentation. It is wrong to “label” a case as Hy’s law based only on serum chemistry findings.

5.3.8 Test Validation and Qualification

Considerable recent thought has been devoted to how best to verify, validate, and qualify new tests and biomarkers for use in detecting and diagnosing
findings of interest such as drug-induced liver injury and many other problems. These words all refer to processes or acts that may be done to establish the test methods as reliable and true, conceptually well related to the problem being evaluated and applicable in clinical use, but their definitions and meanings are often confused. This thinking has been captured in two publications from the US Food and Drug Administration [43,44]. Not included, however, appears to have been the idea of how important specificity is, or the importance of considering the incidence or prevalence when assessing the value of a positive test result validated on common findings when used to screen for rare findings (see Section 5.2.6).

5.4
Closing Considerations

The bulk of this book is devoted to many erudite chapters by basic scientists who focus on new drug development, and who have been very successful in recent years of avoiding or terminating new drugs that cause too much liver injury in comparison with their benefits.

The raising of awareness among FDA reviewers, industry scientists, and consultants to both from academic venues has also been effective in that no drug has been approved by the FDA since the end of 1997 that later had to be withdrawn because of hepatotoxicity.

5.4.1
A Handful of “Do Nots”

Shown earlier in italics are five warnings recently presented at the same conference and emphasized again here as terms that we believe should not be used, inviting comments in disagreement or opposition. The contexts are listed for the five points:

(Section 5.2.3): . . . regulation of serum aminotransferase activity is not a liver function, so activities of biomarker enzymes in serum should not be called “liver function tests.”

(Section 5.2.6): It is unfortunate that Vecchio used the word “predictive” for a single test result, without knowing what the results were before or might be afterward.

(Section 5.3.2): It is wrong to assess severity of liver dysfunction by grading elevations of serum ALT values.

(Section 5.3.6): ROC curves are the same for both common and rare events; they should not be used to validate new tests if neither incidence nor prevalence is considered.

(Section 5.3.7): It is wrong to “label” a case as Hy’s law based only on serum chemistry findings.
These “do nots” are offered as examples of fairly frequent misuse of terms, not just to be semantically pure or as examples of nitpicking, but as instances where the terms we use reflect our sometimes less-than-careful thinking. We accept that objecting to this type of terminology may be disputed, and welcome debate on the issues raised.

5.4.2 Need to Standardize ALT Measurement and Interpretation of Normal Ranges

It is now more than 60 years since the then-medical-student Arthur Karmen announced [45] his discovery of a rapid new spectrophotometric assay of serum activity of the enzyme glutamic–oxaloacetic transaminase (SGOT, now called aspartate aminotransferase) released into the circulation from ischemically injured myocardial tissue. Full details of his method followed in January 1955 [46], which led to a revolution in clinical chemistry by reducing the time needed to obtain results from days of tedious paper chromatography to 5 min in a spectrophotometer, and many awards for Dr. Karmen. His mentors, Drs. John LaDue and Felix Wróbleski, quickly used the method to show that serum glutamic–pyruvic transaminase (SGPT, now called alanine aminotransferase) was useful in diagnosing liver injury to hepatocytes [47]. Despite passage of six decades, we still do not have any international standard for how to measure the activity of serum ALT, nor how best to define the range of normal. An early step in advancing research in the field would be to call together the influential policy makers to agree on exactly how it should be measured, and what should be its range of normal.

5.4.3 Research Opportunities

The power, usefulness, and value of the concepts embodied in the eDISH program now have become evident, and it is time to consider development of “eDISH2” to improve any deficiencies and increase the power of the program to learn more about liver disease. We envision two immediate improvements: (1) ability to reset lines separating nearly normal from elevated values for both ALT and TBL, and (2) ability to use the pretreatment and screening values for measurement of both to establish individual baselines, for comparison with later values after drug treatment is started. The programs do not contain data but can process data, including confidential data in FDA files after submissions for approval and company data developed by expenditures for clinical trials. Previous efforts have focused on points (subjects) in the right upper quadrant, as cases to be investigated further that may represent Hy’s law cases. New work could well be undertaken to examine the other three quadrants for insights into causes for innocent and reversible ALT elevations, causes for TBL rise without ALT increase, and the large group of normal or near-normal values to obtain detailed descriptions of the type of distributions and reasons therefor. The great
power of the computer to process numbers rapidly, combined with the human ability of quick pattern recognition, has been the mainstay of the eDISH success.

Standardization of how best to measure serum ALT activity and what data should be used to validate the healthy normal range are obvious candidates for appropriate research and international consensus.

Another promising area of research might be application of the insights of the late Rev. Thomas Bayes [48] whose logic has led to an appreciation two-and-a-half centuries later that medical differential diagnosis is perhaps a Bayesian process, and that it might be possible to use the stupendous power of computers to streamline and standardize the now artful skill of diagnostic inquiry. It is an area of interest now [49]. If it were possible to reconcile the conflicting opinions and approaches to diagnosing DILI, debated between adherants of the RUCAM (Roussel Uclaf Causality Assessment Method) developed in France 20–25 years ago [50,51] and the proponents of expert consensus [52], it would help resolve a sticky wicket.

We shall see.

References

13 Grunberg, E. and Schnitzer, R.J. (1952) The effect of hydrazine derivatives of...


reactions to drugs. II. An original model for validation of drug causality assessment methods: case reports with positive rechallenge. *Journal of Clinical Epidemiology*, **46** (11), 1331–1336.