Liver INTERNATIONAL

Liver International ISSN 1478-3223

DRUG-INDUCED LIVER INJURY

Reliability of causality assessment for drug, herbal and dietary supplement hepatotoxicity in the Drug-Induced Liver Injury Network (DILIN)

Paul H. Hayashi[1], Huiman X. Barnhart[2], Robert J. Fontana[3], Naga Chalasani[4], Timothy J. Davern[5], Jayant A. Talwalkar[6], K. Rajender Reddy[7], Andrew A. Stolz[8], Jay H. Hoofnagle[9] and Don C. Rockey[10]

1 Division of Gastroenterology and Hepatology, University of North Carolina, Chapel Hill, NC, USA
2 Duke Clinical Research Institute & Department of Biostatistics, Duke University, Durham, NC, USA
3 Division of Gastroenterology, University of Michigan, Ann Arbor, MI, USA
4 Division of Gastroenterology and Hepatology, Indiana University School of Medicine, Indianapolis, IN, USA
5 Division of Gastroenterology, California Pacific Medical Center, San Francisco, CA, USA
6 Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, MN, USA
7 Division of Gastroenterology, University of Pennsylvania, Philadelphia, PA, USA
8 Division of Gastrointestinal and Liver Diseases, University of Southern California, Los Angeles, CA, USA
9 Liver Disease Research Branch, Division of Digestive Diseases and Nutrition, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), NIH, Bethesda, MD, USA
10 Division of Gastroenterology, University of Texas Southwestern, Dallas, TX, USA

Keywords
adverse reaction – diagnosis – drug-induced liver injury – hepatotoxicity – toxicity

Abbreviations
ALT, alanine aminotransferase; AMA, Anti-mitochondrial antibody; ANA, Anti-nuclear antibody; AP, Alkaline phosphatase; APAP, Acetaminophen; ASMA, anti-smooth muscle antibody; AST, aspartate aminotransferase; DILI, Drug-induced Liver Injury; DILIN, Drug-induced Liver Injury Network; HDS, Herbal and Dietary Supplements; LFTs, Liver Function Tests (AST, ALT, AP, bilirubin); NIDDK, National Institute of Diabetes and Digestive and Kidney Diseases; NIH, National Institutes of Health; RUCAM, Roussel Uclaf Causality Assessment Method; ULN, upper limit of normal.

Correspondence
Paul H. Hayashi, MD, MPH, UNC Liver Program, Division of Gastroenterology & Hepatology, CB# 7584 Burnett-Womack Bldg., Room 8011, Chapel Hill, NC 27599-7584, USA
Tel: 919-966-2516
Fax: 919-966-1700
e-mail: paul_hayashi@med.unc.edu

Received 25 October 2013
Accepted 13 March 2014

DOI:10.1111/liv.12540
Liver Int. 2015; 35: 1623–1632

Abstract

Background & Aims: Because of the lack of objective tests to diagnose drug-induced liver injury (DILI), causality assessment is a matter of debate. Expert opinion is often used in research and industry, but its test–retest reliability is unknown. To determine the test–retest reliability of the expert opinion process used by the Drug-Induced Liver Injury Network (DILIN). *Methods:* Three DILIN hepatologists adjudicate suspected hepatotoxicity cases to one of five categories representing levels of likelihood of DILI. Adjudication is based on retrospective assessment of gathered case data that include prospective follow-up information. One hundred randomly selected DILIN cases were re-assessed using the same processes for initial assessment but by three different reviewers in 92% of cases. *Results:* The median time between assessments was 938 days (range 140–2352). Thirty-one cases involved >1 agent. Weighted kappa statistics for overall case and individual agent category agreement were 0.60 (95% CI: 0.50–0.71) and 0.60 (0.52–0.68) respectively. Overall case adjudications were within one category of each other 93% of the time, while 5% differed by two categories and 2% differed by three categories. Fourteen per cent crossed the 50% threshold of likelihood owing to competing diagnoses or atypical timing between drug exposure and injury. *Conclusions:* The DILIN expert opinion causality assessment method has moderate interobserver reliability but very good agreement within one category. A small but important proportion of cases could not be reliably diagnosed as ≥50% likely to be DILI.

In 2004, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) established the Drug-Induced Liver Injury Network (DILIN) as a multicentre study with aims of improving the understanding of the causes, outcomes and molecular mechanisms of hepatotoxicity owing to medications or herbal and dietary supplements (HDS) (1). The Network of eight centres has enrolled >1200 patients into the Prospective Study. Cases meeting laboratory enrolment criteria are enrolled at the discretion of the investigators at each site and based on their clinical suspicion that such abnormalities are at least possibly owing to DILI. Each case has a clinical narrative and a formatted file of demographics, medications, radiography data, histology (when available) and laboratory values in flow sheet format. Serum, plasma and DNA are also collected.

Because there are no objective diagnostic tests, DILI remains a diagnosis of exclusion that requires adequate clinical, laboratory and imaging data. Scoring algorithms for diagnosis are available (2–4), but retest reliability studies are quite limited. Roussel Uclaf Causality Assessment Method (RUCAM) that was developed under the auspices of the Council for International Organizations of Medical Sciences (CIOMS) is the most widely accepted and validated instrument, yet only one study has examined its retest reliability (5). For research purposes, DILIN uses a standardized procedure for expert opinion consensus to adjudicate the likelihood of DILI that is based in part on RUCAM and yields similar categories of DILI likelihood (6). However, reliability of the DILIN expert opinion process has not been critically assessed. Poor reliability would undermine any future mechanistic studies using serum, plasma or DNA from the DILIN subjects and undermine the use of DILIN cases for development of other more accessible diagnostic algorithms for the clinician. Therefore, the aim of this analysis was to assess the interrater, test–retest reliability of the DILIN consensus opinion process.

Methods

DILIN prospective cohort

The DILIN study has been previously described in detail (6). Patients suspected of having liver injury owing to medications or HDS products were enrolled within 24 weeks of injury onset and then followed up prospectively for 6–24 months depending on the pace and completeness of DILI resolution. Because the enrolment window was 24 weeks, cases enrolled at varying time points in their DILI event. Enrolment criteria were (1) serum alanine aminotransferase (ALT) or aspartate aminotransferase (AST) levels > 5 times the upper limit of the normal (ULN) (or pretreatment baseline if abnormal) on two consecutive occasions, or (2) alkaline phosphatase (AP) levels > twice the ULN (or pretreatment baseline if abnormal) on two consecutive occasions, or (3) total serum bilirubin > 2.5 mg/dl, or international normalized ratio (INR) >1.5 with any elevation in serum ALT, AST or AP. Cases meeting laboratory enrolment criteria are enrolled at the discretion of the investigators at each site and based on their clinical opinion. Case enrolment is not restricted by time of onset because some medications are well known to have very long latencies of even years (e.g. nitrofurantoin). We did not restrict enrolment to those with full follow-up data showing resolution of injury (i.e. dechallenge) because such restriction would hinder the prospective nature of DILIN. We wanted to capture this dechallenge data while under study protocol. Injury pattern was categorized as cholestatic ($R < 2$), mixed (R 2–5) or hepatocellular ($R > 5$) where $R = (\text{ALT/ULN}) \div (\text{AP/ULN})$. Severity level was based on INR, bilirubin, signs of liver failure, need for hospitalization and fatal or transplant outcome as previously described (6). Exclusion criteria included acetaminophen hepatotoxicity, prior liver or bone marrow transplant, alcohol-related liver disease, autoimmune hepatitis or genetic liver disease. Patients with compensated chronic hepatitis B, C or with nonalcoholic fatty liver disease were eligible and enrolled at the discretion of the site investigators. For patients with such background liver disease, reviewers used baseline liver enzyme levels, viral serologies and viral nucleic acid tests to judge the presence of DILI vs. exacerbation of underlying liver disease.

At baseline visit, a detailed history was obtained, and clinical, laboratory and imaging results were extracted from records. As reported previously, a DILIN protocol battery of tests to exclude other causes of liver injury was obtained at enrolment if not extractable from chart records (6). Serum, plasma, urine and DNA specimens were sent to a central repository for future studies. Subjects were followed up for at least 6 months and those with persistent liver abnormalities or signs of chronic liver injury were followed up through 24 months.

Ninety (90%) of the cases had complete documentation of data for all 21 parameters adapted from Agarwal *et al.* as essential data for DILI cases (7). (Appendix figure) Nine (9%) were missing complete data for one parameter each (four cases – incomplete viral serologies, four cases – no documented hepatic imaging, one case – incomplete 'washout' of liver biochemistries); one case (1%) was missing complete data for two parameters, viral serologies and autoimmune markers. This last case was considered unlikely to be DILI in large part owing to lack of these data. Overall, the 90 cases with complete data were considered more likely to be DILI than the 10 cases with incomplete data (median DILIN scores of 2 vs. 3, respectively, $P = 0.04$).

DILIN expert opinion process

DILIN causality assessment has been described in detail (6). The process is the same for medication and HDS hepatotoxicity, as there are little data to suggest that a different causality process is necessary or valid. Each case was adjudicated independently by three hepatologists including the site investigator who enrolled the case. Assessment is based on retrospective review of the case history, laboratory data and prospective follow-up study visits. Each reviewer assigns a causality score corresponding to percentages of DILI likelihood in which 1 = definite (>95% likelihood), 2 = very likely (75–95%), 3 = probable (50–74%), 4 = possible (25–49%) and 5 = unlikely (<25%). Disagreements were identified after the three reviewers submitted their scores. Consensus scores were achieved by electronic mail or conference call discussions. Those cases in which agreement could not be reached by the three reviewers were then voted upon by one member from each DILIN site during monthly conference calls. Final score was assigned by majority vote. For cases involving >1 agent, an overall case score and separate individual scores were determined. For example, the overall case score might be 1 (definite DILI) with one agent scoring a 2 (very likely causal) and the other scoring a 4 (only possibly).

From inception in 2004 to April 16, 2009, adjudication was based on results obtained shortly after enrolment, thus simulating the clinician's task to assess at time of presentation. However, when enrolment occurred within days of onset, data on resolution of injury were sparse. After April 16, 2009, the protocol was changed, so that adjudication was done 6 months after enrolment and follow-up data could be included into the assessment.

Reliability cohort and reassessment

The DILIN Data Coordinating Center chose 100 cases by computer-driven random assignment from the Prospective registry. Two cases involving the interval development of new diagnostic information regarding hepatitis E testing were included, but the HEV data were specifically excluded for reassessment (8). Chosen cases were stratified 1:1 across April 16, 2009. Group A included 49 cases enrolled before April 16, 2009, and Group B, 51 cases enrolled afterwards. Group A cases did not have 6-month data for the initial assessment, but these data were available for the reassessment. Group B cases had 6-month data for both the initial and reassessments. We stratified across these two periods to examine whether reliability is influenced by using follow-up data. For reassessments, 92 cases had three new reviewers. Because of an administrative error, eight cases had one previous reviewer and two new reviewers. No cases were reassessed by the site investigator who enrolled the case. The rationale for excluding the enrolling site investigator from reassessments was to minimize recall bias as the enrolling investigator is often the hepatologist who continues to care for the patient. At least 4 months had to elapse before a case could be selected for reassessment. The process of reassessment was otherwise the same.

Analysis

Standard descriptive statistics were used to characterize this cohort of 100 patients and compared to the remaining 983 cases not selected. The original and re-adjudication scores were compared for the 100 cases using weighted kappa statistics owing to ordinal nature of the categories. Because some cases involved >1 drug or HDS product, the reliability at the individual agent level was also examined. Weighted kappa statistics were also determined for Group A and Group B cases separately. Because the two categories of highest likelihood (>95% likelihood and 75–95%) were similar in clinical and research relevance, reliability was assessed collapsing these two categories into one, thus creating quartiles of per cent likelihood. Score differences were categorized as 1, 2 or >2 scores apart and direction of changes were tallied. Cases with scores crossing the 50% likelihood line (1, 2 or 3 vs. 4 or 5) were re-examined in detail to assess factors that might have led to such uncertainty in DILI diagnosis.

Results

Subjects

Among 1083 DILIN cases adjudicated by June 1, 2011, 49 assessed prior to the use of 6-month follow-up data (Group A) and 51 assessed with the use of 6-month data (Group B) were randomly chosen. These cases were similar to those not chosen across a variety of clinical and demographic variables (Table 1). Seventeen of the 100 had initial agreement without need for any discussion whatsoever. This rate is similar to the 20.1% for the total cohort. Of the 100 cases, 69 involved only one medication or HDS product and 31 involved multiple agents. A total of 138 different agents were implicated. The original causality scores were similar between Groups A and B (Table 2).

Reliability of the DILIN expert opinion causality assessment process

The median time between initial and reassessment was 938 days (range 140–2352). Cross-tabulation of scores between original assessment and reassessment is shown in Table 2a. Weighted kappa statistic for score agreement for overall DILI diagnosis was 0.60, 95% confidence interval (CI) 0.50–0.71. Kappa for the 69 single agent cases and 138 individual agent scores were similar [0.59 (0.46–0.71) and 0.60 (0.52–0.68)]. Score agreement tended to be better for Group B compared with Group A for overall case scores [0.67 (0.38–0.81) vs. 0.53 (0.38–0.68)], single agent cases [0.69 (0.53–0.85) vs. 0.49 (0.31–0.68)] and individual agent scores [0.66 (0.56–0.76) vs. 0.49 (0.34–0.63)], although the differences did not reach statistical significance. Combining

Table 1. Features of DILIN cases randomly selected for reassessment compared with cases not reassessed

	Cases reassessed	Cases not reassessed	
Characteristic	*N* = 100	*N* = 983	*P*-value
Demographics			
Age			
Mean (SD)	50.0 (18.59)	48.8 (17.01)	0.41
Gender			
Female	63 (63.0%)	558 (56.8%)	0.24
Race (self report)			
Caucasian	78/100 (78.0%)	764/973 (78.5%)	
Black	12/100 (12.0%)	113/973 (11.6%)	0.96
Asian	3/100 (3.0%)	36/973 (3.7%)	
Other/Multiracial	7/100 (7.0%)	60/973 (6.2%)	
Body mass index (BMI)			
Number w/BMI	88	917	
Mean (SD)	27.0 (6.89)	27.4 (6.50)	0.39
Liver Injury			
Categorized days from primary drug start to DILI onset			
≤1 week	7/88 (8.0%)	81/813 (10.0%)	
2–4 weeks	29/88 (33.0%)	255/813 (31.4%)	
5–12 weeks	26/88 (29.5%)	272/813 (33.5%)	0.84
13–24 weeks	11/88 (12.5%)	85/813 (10.5%)	
>24 weeks	15/88 (17.0%)	120/813 (14.8%)	
No. of concomitant drugs in the 2 months prior to DILI onset			
0–2	24/93 (25.8%)	191/824 (23.2%)	
3–5	25/93 (26.9%)	223/824 (27.1%)	0.83
>5	44/93 (47.3%)	410/824 (49.8%)	
Pattern of liver injury at onset or earliest after onset*			
Cholestatic (R < 2)	25/100 (25.0%)	248/974 (25.5%)	
Mixed (R 2–5)	17/100 (17.0%)	207/974 (21.3%)	0.59
Hepatocellular (R > 5)	58/100 (58.0%)	519/974 (53.3%)	
Peak values between DILI onset and 6 months after enrolment			
ALT (U/L)			
Mean (SD)	1118 (1887)	998 (1481)	0.90
Alkaline Phosphatase (U/L)			
Mean (SD)	390 (381)	410(418)	0.33
Total Bilirubin (mg/dl)			
Median (25th, 75th)	9.0 (1.7, 22.9)	9.7 (2.7, 19.5)	0.94
INR			
Mean (SD)	2.1 (2.49)	1.6 (1.41)	0.07
Adjudication			
Overall causality score			
Definite Greater than 95%	23/100 (23.0%)	173/744 (23.3%)	
Very likely 75–95%	39/100 (39.0%)	312/744 (41.9%)	
Probable 50–75%	19/100 (19.0%)	141/744 (19.0%)	0.22
Possible 25–50%	10/100 (10.0%)	91/744 (12.2%)	
Unlikely Less than 25%	9/100 (9.0%)	27/744 (3.6%)	
Outcomes			
DILIN severity score(6)			
Mild	23/100 (23.0%)	185/744 (24.9%)	
Moderate	19/100 (19.0%)	156/744 (21.0%)	
Moderate-hospitalized	35/100 (35.0%)	221/744 (29.7%)	0.36
Severe	12/100 (12.0%)	129/744 (17.3%)	
Fatal	11/100 (11.0%)	53/744 (7.1%)	

*R = (ALT/ULN) ÷ (AP/ULN).

score categories 1 and 2 into one category did not change the kappa score for overall case score [0.60 (0.48–0.72)], but the kappa for Group B increased to 0.73 (0.59–0.87), while the kappa for Group A fell to 0.44 (0.27–0.61) (Table 2b). Excluding the eight cases that had one repeat reviewer on reassessment did not change the kappa scores significantly (data not shown).

Table 2. Original and reassessment score frequencies and kappa statistics for Groups A, B and total cohort. (a) 5-point likelihood scale (b) 4-point likelihood scale, combining categories 1 and 2

			Reassessment score frequencies						
(a)			1	2	3	4	5	Totals	Kappa (95% CI)
Original score frequencies	Group A *n* = 49	1	9	3	0	1	0	13	0.53 (0.38–0.88)
		2	0	12	6	1	1	20	
		3	0	3	1	3	1	8	
		4	0	0	1	3	1	5	
		5	0	0	0	3	0	3	
	Group B *n* = 51	1	6	3	1	0	0	10	0.67 (0.53–0.81)
		2	4	10	4	1	0	19	
		3	1	2	4	4	0	11	
		4	0	0	1	4	0	5	
		5	0	0	0	0	6	6	
	Groups A & B *n* = 100	1	15	6	1	1	0	23	0.60 (0.50–0.71)
		2	4	22	10	2	1	39	
		3	1	5	5	7	1	19	
		4	0	0	2	7	1	10	
		5	0	0	0	3	6	9	

(b)			1 or 2	3	4	5	Totals	Kappa (95% CI)
Original score frequencies	Group A *n* = 49	1 or 2	24	6	2	1	33	0.44 (0.27–0.61)
		3	3	1	3	1	8	
		4	0	1	3	1	5	
		5	0	0	3	0	3	
	Group B *n* = 51	1 or 2	23	5	1	0	29	0.73 (0.58–0.87)
		3	3	4	4	0	11	
		4	0	1	4	0	5	
		5	0	0	0	6	6	
	Groups A & B *n* = 100	1 or 2	47	11	3	1	62	0.6 (0.48–0.72)
		3	6	5	7	1	19	
		4	0	2	7	1	10	
		5	0	0	3	6	9	

1 = definite (>95% likelihood), 2 = very likely (75–95%), 3 = probable (50–74%), 4 = possible (25–49%), 5 = unlikely (<25%).

Magnitude of disagreement

The magnitude of disagreement on reassessment was small. Ninety three percent of overall scores were the same or differed by only one point (92% for Group A; 94% for Group B). (Fig. 1a). When scores 1 and 2 were combined into one category, 95 of 100 were within one score (92% for Group A; 98% for Group B). (Fig. 1b).

Direction of disagreement

Overall causality reassessment scores had a lower likelihood of DILI compared with the initial evaluation scores, with a reduction in the average initial score – average reassessment score of −0.22 (range from 3 to −2). Of the 45 cases where scores differed, 15 were considered more likely to be DILI, while 30 were considered less likely to be DILI on reassessment. Direction of changes was similar between groups (Group A: 7 more likely, 17 less likely; Group B: 8 more likely, 13 less likely; *P* = ns).

Because there were more cases originally scoring 1–2 (62%) than 4–5 (19%) reflecting the careful consideration of other potential causes for liver injury before enrolment, the DILIN cohort was prone to a ceiling effect. Therefore, we looked at cases originally scoring in the middle as well as proportionate increases vs. decreases in the reassessment scoring of cases originally scoring a 2 or 4 respectively. Cases initially scored as a 3 still tended to score *less* likely to be a DILI event (42% vs. 32%), and cases originally scoring 2 were reassessed as *less* likely to be DILI proportionately more often than cases scoring 4 were reassessed as *more* likely (33% vs. 20%). (Table 3).

Cases crossing the 50% likelihood on reassessment

There were 14 (14%) cases with scores that crossed the 50% likelihood threshold on reassessment (scores 1–3

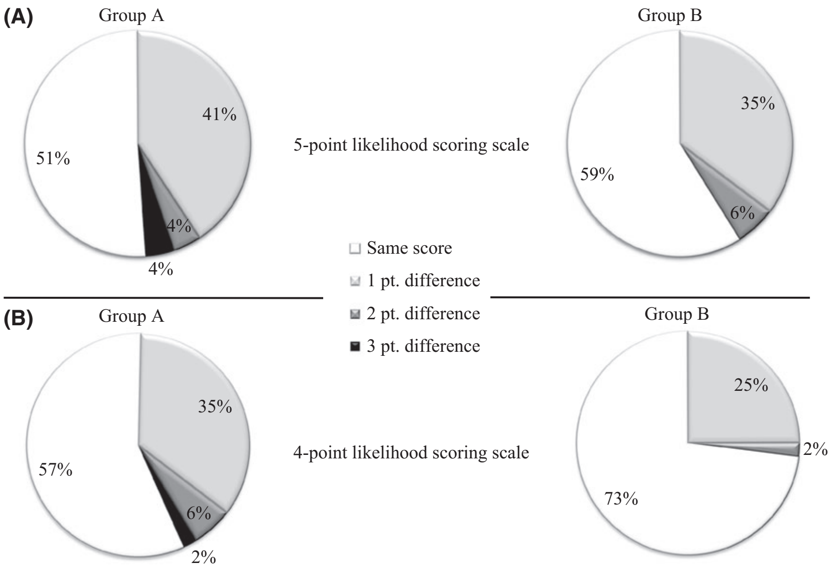

Fig. 1. Magnitude of differences between original and reassessment scores expressed as percentages of cases with 0, 1, 2 and 3 point differences. There were no cases differing by four points. Groups A and B shown separately. (A) Percentages for the 5-point likelihood scoring scale. (B). Percentages for a 4-point likelihood scoring scale that combines scores 1 and 2.

Table 3. Direction of score changes on reassessment stratified by original score

Original score	*N*	Less likely DILI	More likely DILI	No change
1	23	35%	NA*	65%
2	39	33%	10%	57%
3	19	42%	32%	26%
4	10	10%	20%	70%
5	9	NA*	33%	67%

*NA = not applicable.

vs. 4–5). Eight cases were from Group A and 6 from Group B (*P* = ns). In these 14 cases, 12 scores crossed below the 50% threshold, while only two crossed the line towards more likely DILI (Table 4). Nine of 14 (64%) differed by one point only (i.e. between 3 and 4), but five differed by >1 point. Eleven involved only one drug or HDS product. No implicated agents appeared more than once in these 14 cases. Although not statistically significant (*P* = 0.51), the reassessment score differed from the site investigator's initial assessment score more often (12 of 14) than it did for the other two initial reviewers (9 of 14 for both).

Review of these 14 cases and recorded comments by reviewers suggest two major reasons for diagnostic uncertainty. Four cases had uncertain or inconsistent timing between agent exposure and liver injury; seven had competing diagnoses and one had both reasons (Table 4).

Discussion

The DILIN expert opinion process for causality assessment has moderate test–retest reliability on a 5-point scale of likelihood, but agreement within one category of likelihood was very good at 93% of cases. Inclusion of 6-month follow-up data tended to improve concordance, suggesting that diagnostic reliability probably improves when longer follow-up is available. These data are critical in establishing the DILIN registry as a source of tissue and serum for mechanistic studies. These data also help establish the registry as a reliable source of cases for the development of diagnostic instruments that are clinically accessible. While consensus expert opinion may be a reliable diagnostic method, it is cumbersome, inaccessible to the clinician and used only for research purposes. Development of a computerized diagnostic tool for clinicians will need large registries of reliably diagnosed cases (9).

The reliability of the DILIN consensus process is better than that reported for individual assessments without consensus. Studies of reliability between individuals diagnosing adverse drug reactions of all types yield kappa statistics as low as 0.05–0.2. The consensus process used by DILIN substantially elevates the interrater reliability (kappa 0.60). The consensus of three reviewers per case attenuates individual biases and variations in experience. Moreover, the collective experience and expertise of the larger causality committee (15–25 hepatologists per call) is brought to bear on cases where the three reviewers cannot agree. Thus, the consensus process elevates the test–retest reliability to a level comparable to that reported for histological liver diagnoses that also rely on consensus expert opinion. Chronic viral hepatitis biopsies reviewed by four expert liver histopathologists produced kappa statistics for disease activity and fibrosis of 0.43 and 0.59 respectively (10). Similar reliability results were observed for the histological

Table 4. Agents, scores and causes for DILI diagnosis uncertainty in cases crossing the 50% likelihood threshold on reassessment

No.	Agents	Initial score	Reassessment score	Potential reasons for DILI diagnosis uncertainty	
1	Ranitidine	2	4	Timing Competing diagnosis	Short latency (2 days); choledocholithiasis
2	Duloxetine	1	4	Competing diagnosis	Biliary obstruction
3	Topiramate	2	5	Competing diagnosis	Hepatitis C, acute
4	Cefuroxime/Nystatin	3	4	Competing diagnosis	Autoimmune hepatitis
5	Cephalexin/Levofloxacin	3	5	Competing diagnosis	Hepatitis C, chronic
6	Experimental agent	4	3	Competing diagnosis	Ischaemic hepatopathy
7	Ezetimibe/Simvastatin	3	4	Competing diagnosis	Occult alcohol
8	Sulfamethoxazole-Trimethoprim	2	4	Timing	Unclear timing from drug exposure to elevated liver enzymes
9	Ceftriaxone/Ampicillin-Sulbactam/ Fluconazole	3	4	Competing diagnosis	Ischaemic hepatopathy; occult acetaminophen overdose
10	Ciprofloxacin	4	3	Competing diagnosis	Hepatitis, atypical viral (giant cell)
11	Hydroxycut$_{TM}$ (HDS*)	3	4	Timing	Long latency (approximately 75 days) after stopping agent
12	Azithromycin	3	4	Competing diagnosis	Ischaemic hepatopathy; occult acetaminophen overdose
13	Ultra Vitality$_{TM}$ (HDS*)	3	4	Timing	Long latency (approximately 2190 days) while on agent
14	6-Mercaptopurine	3	4	Timing	Unclear timing from drug exposure to elevated liver enzymes

*Herbal Dietary Supplement.

diagnosis of NASH (kappa of 0.61) (11). Therefore, if DILI diagnosis by expert opinion is to be considered as reliable as interpretation of liver biopsies, our data suggest that a rigorous consensus process must be incorporated.

Reliability measured by weighted kappa and per cent agreement within one score were consistently better in Group B for overall case and individual agent scores, although the differences did not reach statistical significance probably as a result of small sample size. Group A's lack of requirement of 6-month follow-up data is a significant handicap as most medical diagnoses become more reliable over time. In addition, operational variability probably decreased over time as the DILIN reviewers became more familiar with the scoring scale and consensus process. For these reasons, Group B more accurately reflects the DILIN's current reliability. When categories 1 and 2 were combined, the kappa for Group B improved to 0.73 with 98% of case reassessments being within one score of each other.

Cases originally scoring in the middle of the scale at 3 were most likely to have different scores on reassessment (Table 4) probably as a result of less certainty of DILI diagnosis as well as being midway in a 5-point scale. Reassessment scoring tended to decrease as opposed to increase in likelihood of DILI (Table 4). Exclusion of the enrolling site investigator from reassessment may have contributed to this finding, because the site investigator was often the hepatologist who also provided clinical care to the patient. Such first-hand knowledge of the case may provide more accurate causality assessment and thereby provide the site investigator a stronger position from which to advocate for a particular score during discussions. Nevertheless, the consensus process is robust enough to still produce reasonable kappa values and very close agreement within one score without first-hand knowledge of the cases on reassessment.

As with any diagnostic tool, DILIN expert opinion is prone to interval discoveries. Recently, a small percentage of DILI cases were discovered to have evidence of acute hepatitis E virus (HEV) infection. (8, 12). Two patients included in this study had data on hepatitis E infection discovered between the two assessments. For studying retest reliability of the DILIN *processes* only, we expunged this HEV data from the documents reviewed by the second set of reviewers. However, accuracy mandates that such interval discoveries be continually incorporated and cases reassessed as were done for the DILIN HEV cases (8). This reassessment process and updating of scores is built into the DILIN protocol. Concerns over such changes should not halt efforts to accrue cases and model diagnostic instruments using the DILIN registry for training and validation purposes. Ideally, any new diagnostic models for clinicians should be malleable enough to incorporate new discoveries as they become available.

DILIN cases vary in complexity particularly with regard to whether or not competing causes of liver injury are identified. The 14 (14%) cases that could not be reliably diagnosed as DILI with at least 50% certainty (i.e. cases crossing between 3 and 4 on reassessment) were some of the more complicated cases that had equivocal presenting diagnostic or longitudinal data. Because these complex cases are a part of clinical practice, a detailed examination of them may guide the building of an accessible diagnostic instrument for clinicians. None of these 14 cases involved isoniazid (INH) or amoxicillin-clavulanate, although these drugs were the two most frequently implicated in the DILIN registry accounting for >15% of all cases (13). The well-established signature patterns of injury for these agents (14–16) probably make DILI easier to adjudicate on one side or the other of the 50% threshold. Giving more weight to such signature presentations may improve the reliability of future diagnostic algorithms.

An alternative cause of liver injury was identified in 10 of the 14 cases and many of these alternative diagnoses lack objective, confirmatory tests (e.g. ischaemic hepatitis, alcohol-related liver injury). Alternative diagnoses are well known to complicate attribution of a liver injury to a specific medication or herbal product (17, 18). Incorporation of diagnostic criteria for competing disorders (e.g. International Autoimmune Hepatitis Group diagnostic criteria) (19) in expert opinion and future causality instruments may improve reliability in such cases. Expert opinion struggled with missing of precise data on timing of liver enzyme abnormalities and agent exposure in five cases. Perhaps, clear-cut timing of agent start and stop and enzyme elevation should be a minimum requirement for assessment akin to minimum requirements suggested for DILI case reporting (20). Rare or unknown hepatotoxicity may have contributed to uncertainty in two cases (ranitidine and an experimental agent), so it would be useful to have standardized scoring of published data that is more precise than what is found in RUCAM. The LiverTox on-line textbook developed by the NIDDK, National Library of Medicine and DILIN contains an extensive listing of publications related to DILI attributed to several hundred agents (http://livertox.nih.gov/index.html). This site may prove useful in standardizing assessment of published data on a particular agent.

Progress in the prevention, early detection and treatment of DILI will require well-characterized and prospectively followed cases of injury attributed to specific agents. Only with large registries of reliably diagnosed DILI will progress be made in determining the molecular mechanisms of DILI. This analysis suggests that the expert opinion causality assessment process used in DILIN will provide a cohort in whom the majority of cases are reliably diagnosed as DILI or not. Specifically, the updated and now well-practised causality assessment process yields moderate diagnostic reliability based on kappa statistics and excellent agreement within one score.

While such internal reliability is encouraging, external reliability among non-DILIN experts using the same consensus process would be worth examining. The DILIN experience is limited to the USA, and therefore applicability to other countries remains unclear. The DILIN expert opinion process also lacks quantitative scores of individual characteristics that may always hinder retest reliability. Importantly, our study does not address the daunting problem of validity, as there is no gold standard for the diagnosis of DILI. Some have suggested using the RUCAM followed by expert opinion in a two-step process to enhance reliability and validity (21).

Finally, a minority of cases straddles the 50% likelihood line and eludes a reliable diagnosis of DILI vs. not DILI usually because of lack of typical presentation, imprecise information on timing or the presence of competing causes of liver injury. Cases scoring in this middle range in general, but particularly when concerns over timing or competing diagnoses are raised, will need to be reviewed carefully if used for mechanistic studies and deserve special attention if a more automated and widely accessible causality assessment instrument is to be developed.

Acknowledgements

We recognize all the site investigators, co-investigators, study coordinators and staff for their invaluable work on this study and on the DILIN. Please see the DILIN website at https://dilin.dcri.duke.edu/publications-1 for a complete listing of DILIN personnel.

Financial support: The DILIN Network is structured as a U01 cooperative agreement with funds provided by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) under grants: 2U01-DK065176-06 (Duke), 2U01-DK065201-06 (UNC), 2U01-DK065184-06 (Michigan), 2U01-DK065211-06 (Indiana), 5U01DK065 193-04 (UConn), 5U01-DK065238-08 (UCSF/CPMC), 1U01-DK083023-01 (UTSW), 1U01-DK083027-01 (TJH/UPenn), 1U01-DK082992-01 (Mayo), 1U01-DK083020-01 (USC). Additional funding is provided by CTSA grants: UL1 RR025761 (Indiana), UL1 RR025747 (UNC), UL1 RR024134 (UPenn), UL1 RR024986 (Michigan), UL1 RR024982 (UTSW), UL1 RR024150 (Mayo) and in part by the Intramural Research Program of the NIH, National Cancer Institute.

Conflict of interest: RJ Fontana: Consultant GSK, Tibotec, Vertex, Gilead; N Chalasani: Consultant Merck, Abbvie, Aegerion, Salix, BMS, Lilly; KR Reddy: Advisory Board BMS, Abbvie, Gilead, Merck Genetech-Roche, Vertes, Jannsen. The rest of the authors do not have any disclosures to report.

References

1. Hoofnagle JH. Drug-induced liver injury network (DILIN). *Hepatology* 2004; **40**: 773.
2. 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**: 1323–30.
3. 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**: 1331–6.

4. Maria VA, Victorino RM. Development and validation of a clinical scale for the diagnosis of drug-induced hepatitis. *Hepatology* 1997; **26**: 664–9.
5. Rochon J, Protiva P, Seeff LB, *et al.* Reliability of the Roussel Uclaf causality assessment method for assessing causality in drug-induced liver injury. *Hepatology* 2008; **48**: 1175–83.
6. Fontana RJ, Watkins PB, Bonkovsky HL, *et al.* Drug-Induced Liver Injury Network (DILIN) prospective study: rationale, design and conduct. *Drug Saf* 2009; **32**: 55–68.
7. Agarwal VK, McHutchison JG, Hoofnagle JH. Important elements for the diagnosis of drug-induced liver injury. *Clin Gastroenterol Hepatol* 2010; **8**: 463–70.
8. Davern TJ, Chalasani N, Fontana RJ, *et al.* Acute hepatitis E infection accounts for some cases of suspected drug-induced liver injury. *Gastroenterology* 2011; **141**: 1665–72. e1-9.
9. Garcia-Cortes M, Stephens C, Lucena MI, Fernandez-Castaner A, Andrade RJ. Causality assessment methods in drug induced liver injury: strengths and weaknesses. *J Hepatol* 2011; **55**: 683–91.
10. Rousselet MC, Michalak S, Dupre F, *et al.* Sources of variability in histological scoring of chronic viral hepatitis. *Hepatology* 2005; **41**: 257–64.
11. Kleiner DE, Brunt EM, Van Natta M, *et al.* Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology* 2005; **41**: 1313–21.
12. Dalton HR, Fellows HJ, Stableforth W, *et al.* The role of hepatitis E virus testing in drug-induced liver injury. *Aliment Pharmacol Ther* 2007; **26**: 1429–35.
13. Chalasani N, Fontana RJ, Bonkovsky HL, *et al.* Causes, clinical features, and outcomes from a prospective study of drug-induced liver injury in the United States. *Gastroenterology* 2008; **135**: 1924–34, 34 e1-4.
14. Zimmerman HJ. Antituberculosis agents. In: Zimmerman HJ, ed. *Hepatotoxicity: The Adverse Effects of Drugs and Other Chemicals on the Liver*, 2nd edn. Philadelphia: Lipponcott Williams & Wilkins, 1999; 611–21.
15. Lucena MI, Andrade RJ, Fernandez MC, *et al.* Determinants of the clinical expression of amoxicillin-clavulanate hepatotoxicity: a prospective series from Spain. *Hepatology* 2006; **44**: 850–6.
16. Verma S, Kaplowitz N. Hepatotoxicity of antituberculosis drugs. In: Kaplowitz N, DeLeve LD, eds. *Drug-Induced Liver Disease*, 2nd edn. New York: Informa Healthcare USA, 2007; 547–66.
17. Aithal GP, Rawlins MD, Day CP. Accuracy of hepatic adverse drug reaction reporting in one English health region. *BMJ* 1999; **319**: 1541.
18. Teschke R, Schulze J, Schwarzenboeck A, Eickhoff A, Frenzel C. Herbal hepatotoxicity: suspected cases assessed for alternative causes. *Eur J Gastroenterol Hepatol* 2013; **25**: 1093–8.
19. Hennes EM, Zeniya M, Czaja AJ, *et al.* Simplified criteria for the diagnosis of autoimmune hepatitis. *Hepatology* 2008; **48**: 169–76.
20. Agarwal VK, McHutchison JG, Hoofnagle JH. Important elements for the diagnosis of Drug-Induced Liver Injury. *Clin Gastroenterol Hepatol* 2010; **8**: 463–70.
21. Teschke R, Eickhoff A, Schulze J. Drug- and herb-induced liver injury in clinical and translational hepatology: causality assessment methods, quo vadis? *J Clin Translat Hepatol* 2013; **1**: 59–74.

Appendix

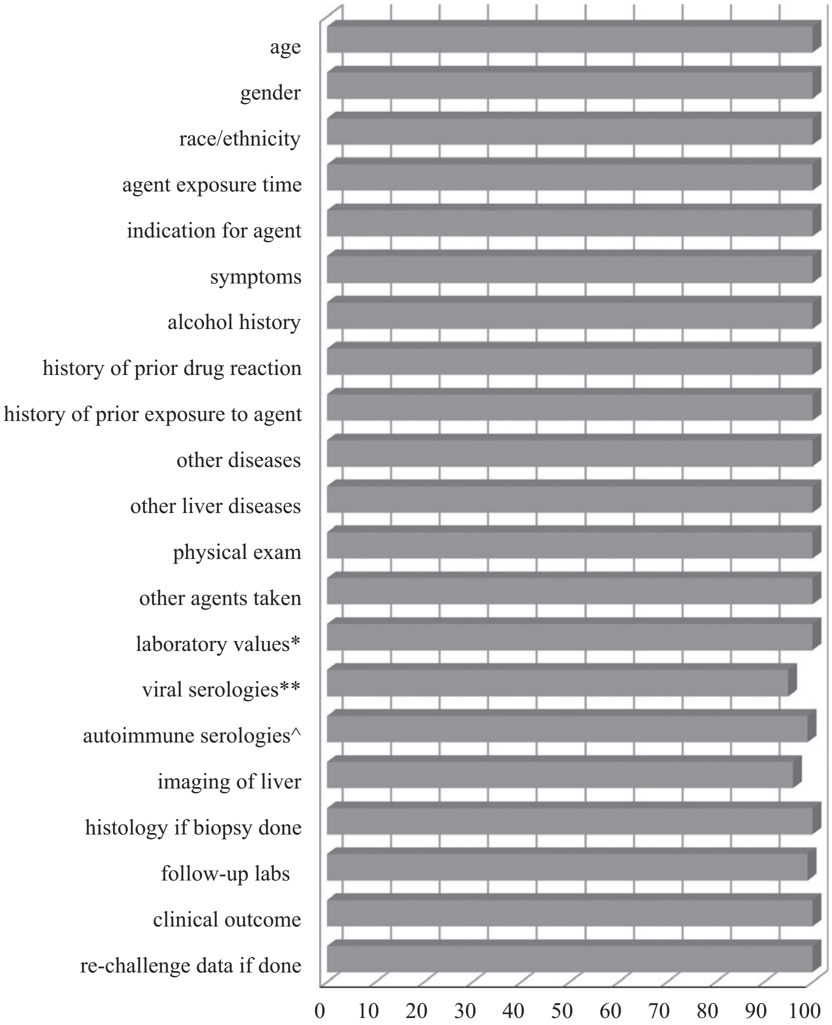

Figure. Completeness of data at adjudication for parameters adapted from Agarwal *et al.* (7).
*Laboratory values including liver biochemistries, INR, cell count at onset and enrollment.
**Complete viral serologies for acute hepatitis A, B and C.
^Autoimmune serologies (ANA, ASMA).