Reliability of the Roussel Uclaf Causality Assessment Method for Assessing Causality in Drug-Induced Liver Injury*

James Rochon,[1] Petr Protiva,[2] Leonard B. Seeff,[3] Robert J. Fontana,[4] Suthat Liangpunsakul,[5] Paul B. Watkins,[6] Timothy Davern,[7] and John G. McHutchison,[1] for the Drug-Induced Liver Injury Network (DILIN)

The Roussel Uclaf Causality Assessment Method (RUCAM) was developed to quantify the strength of association between a liver injury and the medication implicated as causing the injury. However, its reliability in a research setting has never been fully explored. The aim of this study was to determine test-retest and interrater reliabilities of RUCAM in retrospectively-identified cases of drug induced liver injury. The Drug-Induced Liver Injury Network is enrolling well-defined cases of hepatotoxicity caused by isoniazid, phenytoin, clavulanate/amoxicillin, or valproate occurring since 1994. Each case was adjudicated by three reviewers working independently; after an interval of at least 5 months, cases were readjudicated by the same reviewers. A total of 40 drug-induced liver injury cases were enrolled including individuals treated with isoniazid (nine), phenytoin (five), clavulanate/amoxicillin (15), and valproate (11). Mean ± standard deviation age at protocol-defined onset was 44.8 ± 19.5 years; patients were 68% female and 78% Caucasian. Cases were classified as hepatocellular (44%), mixed (28%), or cholestatic (28%). Test-retest differences ranged from −7 to +8 with complete agreement in only 26% of cases. On average, the maximum absolute difference among the three reviewers was 3.1 on the first adjudication and 2.7 on the second, although much of this variability could be attributed to differences between the enrolling investigator and the external reviewers. The test-retest reliability by the same assessors was 0.54 (upper 95% confidence limit = 0.77); the interrater reliability was 0.45 (upper 95% confidence limit = 0.58). Categorizing the RUCAM to a five-category scale improved these reliabilities but only marginally. *Conclusion:* The mediocre reliability of the RUCAM is problematic for future studies of drug-induced liver injury. Alternative methods, including modifying the RUCAM, developing drug-specific instruments, or causality assessment based on expert opinion, may be more appropriate. (HEPATOLOGY 2008;48: 1175-1183.)

Abbreviations: CRF, case report form; DILI, drug-induced liver injury; DILIN, Drug-Induced Liver Injury Network; ILIAD, idiosyncratic liver injury associated with drugs; MAD, maximum absolute difference; NIDDK, National Institute of Diabetes and Digestive and Kidney Diseases; PI, principal investigator; RUCAM, Roussel Uclaf causality assessment method; SD, standard deviation; U95CL, upper 95% confidence limit

From the [1]Duke Clinical Research Institute, Duke University, Durham, NC; [2]University of Connecticut Health Sciences Center, University of Connecticut, Farmington, CT; [3]National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD; [4]University of Michigan, Ann Arbor, MI; [5]Indiana University School of Medicine, Indianapolis, IN; [6]University of North Carolina at Chapel Hill, Chapel Hill, NC; and [7]University of California at San Francisco, San Francisco, CA.

Received November 26, 2007; accepted May 20, 2008.

Supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) under grants 1U01DK065201, 1U01DK065193, 1U01DK065184, 1U01DK065211, 1U01DK065238, and 1U01DK065176.

**This is publication #1 from the Drug-Induced Liver Injury Network (DILIN).*

Address reprint requests to: James Rochon, PhD, Duke Clinical Research Institute, P.O. Box 17969, Durham, NC 27713. E-mail: james.rochon@duke.edu; fax: 919-668-7055.

Published online in Wiley InterScience (www.interscience.wiley.com).

DOI 10.1002/hep.22442

Potential conflict of interest: Nothing to report.

Identifying drug-induced liver injury (DILI) is a major clinical challenge.[1] Unlike many other medical conditions, no single test or biochemical signal exists to establish a definitive diagnosis. This diagnostic dilemma is heightened by the fact that DILI can mimic virtually all known forms of acute and chronic liver disease. Any confirmation of suspected DILI requires that all the other plausible causes of liver disease be excluded; for example: infection with hepatitis viruses A, B, and C; alcoholic and autoimmune hepatitis; or ischemic or congestive hepatopathy. However, attention must also be directed to the pattern of liver test abnormalities, the duration of latency to symptomatic presentation, the presence or absence of features suggestive of immune-mediated hypersensitivity, and the response to drug withdrawal and rechallenge.[2-5] Thus, unlike many other disease areas, DILI is diagnosed primarily by the clinical judgment of the attending physician and is frequently based on "guilt by association."[6]

Two instruments that have been developed to quantify the strength of association between the liver injury and the implicated medication include the Roussel Uclaf Causality Assessment Method (RUCAM)[7,8] and the Maria and Victorino clinical scale.[9] The RUCAM is composed of seven different criteria; including: the time to onset, clinical course, risk factors, concomitant drugs, non-drug-causes, published information on hepatotoxicity, and the response to any re-administration. The RUCAM score ranges from −8 to +14, with higher values signifying a greater degree of association. The RUCAM was developed on an *ad hoc* basis by consensus opinion among hepatotoxicity experts. Although it appears to be superior to the Maria and Victorino scale[10] and is widely used by the pharmaceutical industry, a number of shortcomings have emerged. There are no explicit instructions on how to interpret and score the individual components, the defining criteria are somewhat dated, the scales for the different components are rather arbitrary, and the final score is not intuitive. Despite these shortcomings, the RUCAM has been used to identify DILI events in case studies of prescription drugs,[11-13] herbal medications,[14,15] epidemiological studies,[16-18] clinical trials,[19] and genotyping studies.[20]

The RUCAM was validated using patients who had been rechallenged with the implicated medication.[8] However, its reliability in a clinical research setting has never been fully explored. In this work, we report the results of an empirical investigation of the test-retest and interrater reliability of the RUCAM. Specifically, the following questions were posed: (1) Is there consistency in the RUCAM score when it is repeated over an interval of time? (2) Is there consistency in the RUCAM score across independent reviewers? and (3) What are the test-retest and interrater reliability coefficients of this instrument?

Materials and Methods

The Drug-Induced Liver Injury Network. In 2003, the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) established the Drug-Induced Liver Injury Network (DILIN) to advance understanding and research into DILI.[21] The network is composed of five clinical sites and their affiliates: a data coordinating center, the NIDDK project office, a biosample repository, and a central liver histopathology core. The network is charged with identifying large numbers of well-defined DILI cases and to collect clinical data and biological samples for the study of pathogenesis.

The Idiosyncratic Liver Injury Associated with Drugs Study. One of the DILIN studies is the Idiosyncratic Liver Injury Associated with Drugs (ILIAD) study. ILIAD is a retrospective study enrolling patients who have experienced significant DILI in the recent past. The protocol was approved by the Institutional Review Boards at the participating institutions and is registered at ClinicalTrials.gov. To be eligible, patients must have been alive at the time of the clinic visit; the implicated medication must have been either isoniazid, phenytoin, combination clavulanic acid/amoxicillin, or valproic acid; the date of onset of the qualifying DILI event must have been on or after January 1, 1994; and there must be sufficient documentation that a causality determination can be made. Individuals less than 2 years old at the time of study enrollment were excluded due to blood volume requirements. January 1, 1994 was chosen because it is considered to be the limit at which accurate, relevant, and complete medical records and charts would be available for abstraction, as well as the time when diagnostic testing for acute and chronic hepatitis C virus infection became widely available in the United States. The four drugs were targeted because they cause severe DILI at a sufficiently high rate compared to other drugs, have a characteristic clinical pattern of injury, and are typically administered to reasonably healthy patients not concurrently receiving other drugs more likely to be hepatotoxic. This allows causality to be assessed in a manner relatively uncontaminated by extraneous factors, and might therefore be considered as the "best-case scenario" for assessing the reliability of the RUCAM.

Cases were enrolled in the study if there was sufficient evidence that the liver injury was causally associated with the implicated drug as judged by the clinical experience of the site principal investigator (PI). The qualifying criterion for notable liver dysfunction was a total serum bilirubin level >2.5 mg/dL on at least one occasion for isoniazid, phenytoin, and clavulanic acid/amoxicillin; the criteria for valproate were a compatible symptomatic clin-

ical presentation severe enough to prompt hospitalization, with evidence of liver dysfunction; that is, international normalized ratio >1.5 or alanine aminotransferase >3× the upper limit of normal, and/or characteristic abnormalities seen on liver biopsy. In all instances, the date of "onset" was defined as the first date on or after starting the implicated medication when the corresponding criterion was detected recognizing that the true date of onset is often unknown.

All enrolled patients reviewed and signed an informed consent document. Some information was collected directly from patients using a telephone or personal interview including demographic information, alcohol consumption history, vital status and demographics of all biological parents, siblings, and children, together with a history of prior liver problems. Detailed clinical data were abstracted from all available medical records including medical conditions and illnesses; detailed exposure to the implicated medication as well as other prescription and herbal medications; signs, symptoms and extra-hepatic manifestations during the liver injury; detailed liver-related biochemical tests including serum alanine aminotransferase, aspartate aminotransferase, alkaline phosphatase, bilirubin, international normalized ratio, and prothrombin time; other laboratory assays as available including the complete blood count with differentials, serum albumin, protein and creatinine levels; the results from serological and other assays and any imaging studies; and whether the patient was rechallenged with the DILI medication. Data were recorded on case report forms (CRFs), and once data validation was complete, the case was made ready for the adjudication process.

Causality Adjudication Process. Causality adjudication was performed by a committee of DILIN investigators. The committee met monthly by teleconference and consisted of the PIs and coinvestigators at the five clinical sites as well as members from the data coordinating center; and the NIDDK. For each case, a subset of the CRF containing data directly relevant to the adjudication process (i.e., the "CRF Subset") was extracted from the database and made available to reviewers. Additionally, a two-page clinical narrative was compiled by the site investigator contributing the case. It provided a summary of the clinical/medical history of the case, but focused on supplemental information that was difficult to capture in a CRF. This included a succinct summary of the presentation, laboratory values, diagnostic studies, and the rationale for attributing the event to drug-induced liver injury.

The clinical narrative and CRF subset were provided to three reviewers. The first was the PI at the DILIN clinical site who enrolled the case and completed the clinical narrative. The other two reviewers were selected at random from other committee members. They came from distinct clinical sites and from sites other than that enrolling the case. The three reviewers worked independently of each other and completed the RUCAM instrument. None of the reviewers had used the RUCAM instrument routinely prior to participating in the study, and operating procedures were developed after consulting with one of the RUCAM authors. The total score was derived using published criteria.[7] This methodology allows the reliability among the three reviewers to be evaluated. To evaluate test-retest reliability, the entire process was repeated. That is, after a "washout" period of at least 5 months (mean = 343; range; 150 to 531 days), each case was reviewed by the same three reviewers on a second occasion. Thus, each case was reviewed by three independent reviewers on two separate occasions.

Statistical Methods. Simple descriptive statistics, that is mean ± standard deviation (SD), and frequency distributions, were used to summarize the characteristics of the study population and the liver injuries. They were also applied to the RUCAM score on each occasion and for the difference between them. A Bland-Altman plot[22] was applied to depict differences between the two occasions. That is, for each case and each reviewer, the difference between the two occasions was plotted against their average to determine if test-retest differences were consistent throughout the range of the RUCAM score. To appreciate differences among the three reviewers, the absolute values of the pairwise differences among the three reviewers were derived. Then, the maximum among these pairwise differences, the maximum absolute difference (MAD), was determined. A MAD value of 0 indicates complete agreement among the three reviewers; a value of 1 indicates a case in which two reviewers agreed and the third differed by only 1 point.

A mixed-effects regression model was applied to perform a formal reliability analysis. To account for inherent variability among the cases, the following variables were included as covariates: clinical site, age greater than or equal to 55 years, gender, any alcohol use, hepatocellular versus cholestatic/mixed liver injury, time to onset, and the peak liver serum enzyme values. Random effect terms included patient, reviewer, and the patient × reviewer interaction, and reliability coefficients were derived from the corresponding variance estimates.[23] A reliability coefficient ranges from 0 to 1, with higher values signifying greater reliability. Because interest is focused on whether the reliability achieves a minimum threshold of acceptability, a one-sided upper 95% confidence limit (U95CL) was derived using bootstrapping methods.[24,25] Sample size calculations were performed[26] to test the null hypothesis that the reliability coefficient was greater than 0.65.

Table 1. Characteristics of ILIAD Cases

Characteristics	Mean ± SD or Percent*
Demographics:	
Age at DILI onset (years) ($n = 39$)	44.8 ± 19.5
Age ≥55 years ($n = 39$)	35.9%
Gender female	67.5%
Ethnicity (not Hispanic or Latino) ($n = 39$)	87.2%
Race	
Caucasian	77.5%
African-American	15.0%
Prior history:	
Prior history of a liver problem	27.5%
Prior episode of jaundice	25.0%
Prior reaction to a drug requiring doctor visit	35.0%
BMI at DILI onset (kg/m^2) ($n = 39$)	28.5 ± 8.3
At least one alcoholic drink prior to taking drug	30.0%
Days from drug start to onset ($n = 38$)	156 ± 455
Hospitalized	72.5%
How long was the patient sick?	
A few days	5.0%
One week	2.5%
2-4 weeks	30.0%
≥1 month	62.5%
Prescribed prednisone	15.0%
Selected signs and symptoms:	
Jaundice	72.5%
Nausea	55.0%
Dark urine	50.0%
Vomiting	35.0%
Abdominal pain	32.5%
Extrahepatic manifestations	17.5%
Rechallenged with DILI medication	2.5%
Type of liver injury ($n = 39$):†	
Hepatocellular ($R \geq 5$)	44%
Mixed ($2 < R < 5$)	28%
Cholestatic ($R \leq 2$)	28%
Diagnostic tests:	
Abnormal ultrasound ($n = 26$)	57.7%
Abnormal abdominal CT ($n = 17$)	58.8%
Abnormal abdominal MRI ($n = 2$)	100%
Abnormal biopsy ($n = 14$)	100%
Underwent liver transplantation	15.0%

*Based on the complete sample of 40 DILI cases unless otherwise indicated.
†R is defined as (ALT/ULN) ÷ (AP/ULN) at the time of onset. BMI, body mass index; CT, computed tomography; MRI, magnetic resonance imaging.

Power was set to 80%. Anticipating that the interrater reliability would in fact be 0.80, with three reviewers a minimum sample of 40 participants was required.

Results

Patient Characteristics. The study was conducted with the first 40 patients enrolled in the ILIAD study including nine cases of isoniazid, five cases of phenytoin, 15 cases of clavulanate/amoxicillin, and 11 cases of valproate hepatotoxicity. Table 1 provides their clinical characteristics. Mean (±SD) age at onset was 44.8 ± 19.5 years, ranging from 3.5 to 78.5; of these, 36% were 55 years of age or older. A total of 68% of patients were female; 78% were Caucasian and 15% were African-American. Of 40 patients, 11 (28%) reported a known history of pre-existing liver disease including abnormal liver biochemical tests (three patients), chronic hepatitis C virus infection (two patients), hemochromatosis (one patient), and unspecified cirrhosis (one patient). Mean body mass index at DILI onset was 28.5 kg/m^2, ranging from 16.0 to 51.1. A total of 30% reported having at least one alcoholic beverage during the 1-month interval prior to starting the implicated medication, and most reported 1-2 drinks per occasion (not shown).

Severity of the Liver Injury. The mean (±SD) number of days from drug start to onset of liver injury was 156 ± 455; however, this mean was heavily influenced by a small number of large values. The median time to onset was 42.5 days. Of the 40 patients, 73% were hospitalized, most (63%) were ill for several months, and 15% were prescribed prednisone. During the liver injury, all patients experienced at least one symptom, including jaundice (73%), nausea (55%), dark urine (50%), vomiting (35%), and abdominal pain (33%); 18% had extrahepatic manifestations. Only one of 40 patients (2.5%) was rechallenged. Cases were characterized as being cholestatic (28%), mixed (28%), or hepatocellular (44%) in presentation, based on the R-ratio on the date of onset. Ultrasound of the liver was performed in 26 of 40 cases, and was found to be abnormal in 15 (58%) patients. Ultimately, six patients (15%) required liver transplantation. Mean (±SD) peak serum test results during the injury, by implicated drug, are provided in Table 2, and reflect characteristic patterns of liver injury caused by these drugs.

Site PI versus External Reviewers. Of the 120 scores expected on each occasion, 119 and 116 were available on the first and second occasions, respectively. Missing reviews were due to investigators who left DILIN in the intervening period or were otherwise unavailable. Preliminary analyses revealed a significant difference between the site PI and the external reviewers as a group. After adjusting for covariates, the RUCAM score (mean ± standard error) for the site PI was 7.2 ± 0.5 versus 6.4 ± 0.5 for the external reviewers ($P = 0.007$). We therefore

Table 2. Mean ± SD Peak Serum Tests Observed During the Liver Injury Expressed as a Multiple of the ULN

Serum Test	Isoniazid	Phenytoin	Clavulanate/Amoxicillin	Valproate
AST	48.0 ± 28.5	36.9 ± 48.6	15.6 ± 25.8	16.1 ± 21.4
ALT	32.1 ± 13.3	20.7 ± 13.2	12.5 ± 23.2	14.3 ± 20.1
AP	2.7 ± 1.2	5.9 ± 9.1	4.0 ± 2.1	1.7 ± 1.1
Bilirubin	15.8 ± 9.7	10.9 ± 10.4	14.4 ± 14.6	6.7 ± 10.6

ULN, upper limit of normal; AST, aspartate aminotransferase; ALT, alanine aminotransferase; AP, alkaline phosphatase.

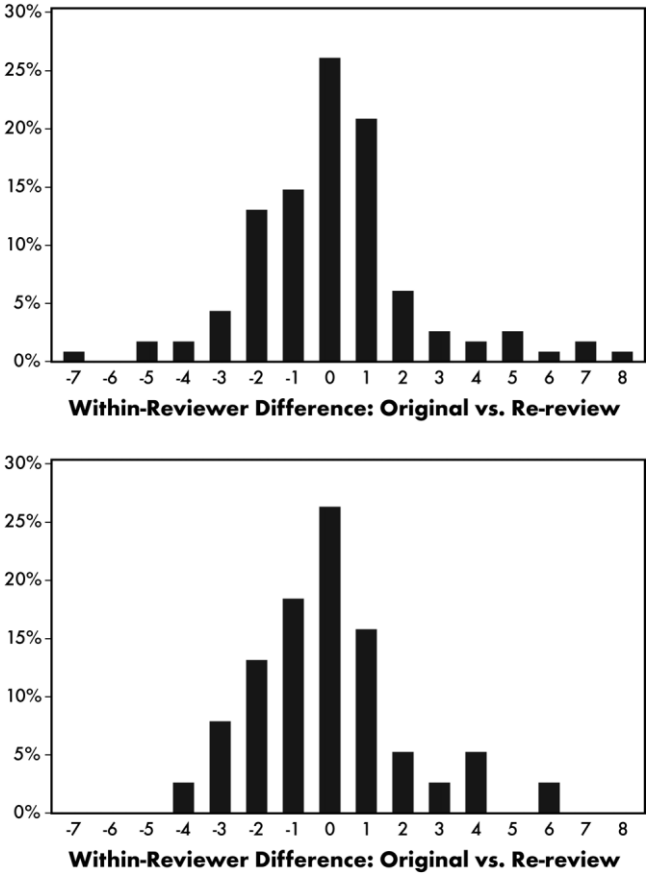

Fig. 1. Histogram of the within-reviewer differences from the first occasion to the second occasion. (A) All reviewers. (B) Site PIs only.

distinguished between these two groups in subsequent analyses.

Comparison Between the Two Occasions. Mean adjusted RUCAM scores (±standard error) on the two occasions were 6.4 ± 0.5 and 6.6 ± 0.5, respectively (P = 0.29). However, this overall result masked considerable variability on a case-by-case basis. The difference between the two occasions pooled across reviewers and patients is shown in Fig. 1A. Differences ranged from −7 to +8 with positive differences largely offset by negative ones. Site PIs were less variable than the external reviewers (Fig. 1B) with their test-retest differences ranging from −4 to +6. Overall, there was complete agreement in only 26% of cases, with differences greater than two or three points in absolute value in 19% and 12% of cases, respectively. A Bland-Altman plot (Fig. 2) revealed that differences were roughly consistent throughout the range of the RUCAM score, with perhaps smaller test-retest differences towards the lower end of the scale.

Comparisons across the Three Reviewers. On average, the MAD among the three reviewers was 3.1 on the first occasion and 2.7 on the second occasion. However, much of this variability could be attributed to differences between the site PI and the external reviewers. Figure 3 summarizes the MAD between the two external reviewers (only) on the two occasions. The MAD ranged from 0 to 7, with average MADs of 2.0 and 1.3 on the first and second occasions, respectively.

Reliability Analysis. From the mixed-effects statistical model, the overall test-retest reliability was 0.54 (U95CL = 0.77). The interrater reliability was 0.45 (U95CL = 0.58). Considering the site PIs only, the test-retest reliability improved to 0.65 (U95CL = 0.84). Among the external reviewers, however, the test-retest reliability was 0.43 (U95CL = 0.77), while the interrater reliability was 0.46 (U95CL = 0.63).

The small samples sizes precluded performing separate analyses for each of the implicated drugs. However, the reliability appeared to be lower for valproic acid. For example, test-retest differences of one point or less in absolute value were observed in 65% of isoniazid reviews, 60% of phenytoin reviews, and 64% of clavulanate/amoxicillin reviews, but in only 55% of those with valproic acid (data not shown). Severity of liver disease may have also played a role. Combining patients with a prior liver injury and/or requiring liver transplant (n = 17) revealed test-retest differences of one point or less in absolute value in 55% of reviews. The corresponding number for less severe cases (n = 23) was 67%. Moreover, of the seven components comprising the RUCAM, Question 2 (time course of the liver injury) and Question 5 (potential nondrug causes) exhibited the greatest test-retest differences (not shown). In particular, Question 5 ranges from −3 to +2 and gave rise to test-retest differences ranging from −5 to +5.

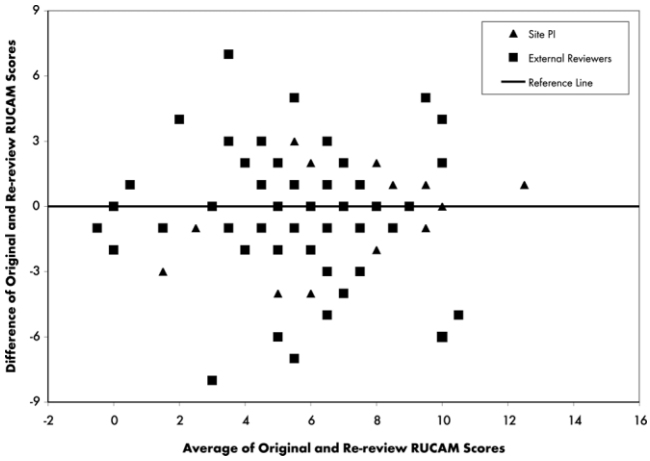

Fig. 2. Bland-Altman plot of test-retest differences versus their mean. The purpose is to determine if the test-retest differences were consistent throughout the range of the RUCAM score. This would be reflected by a constant level of scatter about the reference line. Deviations form this pattern suggests that consistency varies from one place to another in the scale and casts doubt on its overall reliability.

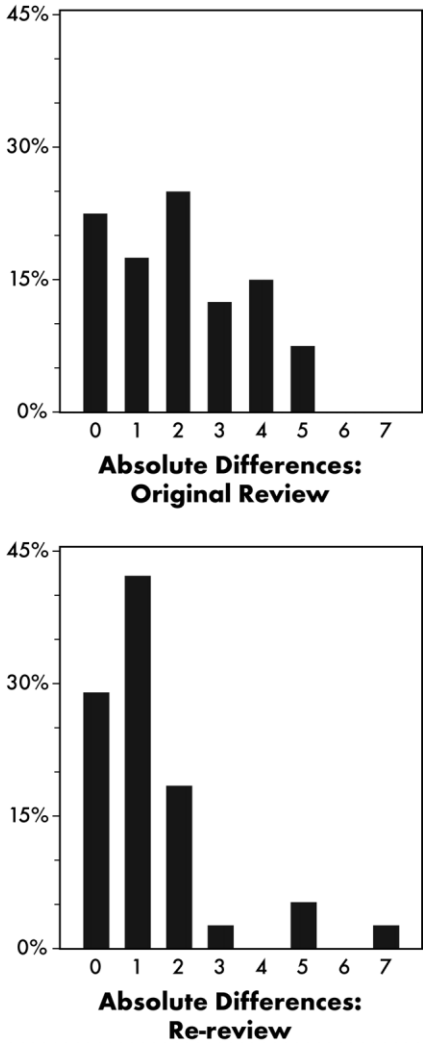

Fig. 3. Histogram of the maximum absolute difference between the two external reviewers. (A) First occasion. (B) Second occasion.

Both questions require the reviewers to interpret complicated clinical information, and they might be more effective if they were split into separate questions.

RUCAM as a Categorized Score. Benichou et al.[8] suggested that the RUCAM could be collapsed into a five-category scale with categories, highly probable (>8), probable (6-8), possible (3-5), unlikely (1-2), and excluded (≤0). Pooling across reviewers and occasions (n = 235), cases were classified as highly probable (15%), probable (53%), possible (24%), unlikely (3%), and excluded (4%). The site PIs again attributed a significantly greater causality category than the external reviewers (P = 0.009). The distribution of the test-retest differences is provided in Table 3, and ranged from −3 to +2. Overall, there was complete agreement in 74 of 115 (64%) of reviews, with a difference of one category or less in absolute value in 91% of reviews. Table 4 provides the MAD between the two external reviewers on each occasion.

Table 3. Distribution of the Between-Occasion Difference in the Five-Category RUCAM Score

	External Reviewers		Site PIs	
Difference	Frequency (n = 77)	Percent	Frequency (n = 38)	Percent
−3	3	3.9	0	0.0
−2	1	1.3	2	5.3
−1	12	15.6	4	10.5
0	49	63.6	25	65.8
1	9	11.7	6	15.8
2	3	3.9	1	2.6

The RUCAM is categorized as follows: highly probable (>8), probable (6-8), possible (3-5), unlikely (1-2), and excluded (≤0).

There was complete agreement (MAD = 0) in only 50% and 63% of cases, respectively; MADs of one point or less were observed in 95% and 90% of cases on the two occasions.

Using the categorized scale, the overall test-retest reliability was 0.51 (U95CL = 0.76); the interrater reliability was 0.34 (U95CL = 0.49). Considering the site PIs only, the test-retest reliability improved to 0.61 (U95CL = 0.81). Among the external reviewers, the test-retest reliability was 0.54 (U95CL = 0.77), while the interrater reliability was 0.54 (U95CL = 0.59).

To gain a better appreciation of the clinical implication of this result, Table 5 presents a cross-tabulation of the Reviewer A score versus the Reviewer B score on the categorized RUCAM pooled across the two occasions. Thus, of the 11 reviews in which Reviewer A scored the DILI case as highly probable, Reviewer B agreed in only five instances. Similarly, of the 21 reviews in which Reviewer B scored the case as possible, Reviewer A agreed in only six instances. Question 2 (time course of the liver injury) and to a lesser extent Question 1 (time to onset) and Question 4 (concomitant drugs) were largely responsible for the greatest interrater disagreements. In a research setting, this table might be collapsed further into a simple 2 × 2 table by combining the first two categories and the last three categories to rule in or rule out a DILI case, respectively. Even when reduced to this most basic level, how-

Table 4. Maximum Absolute Difference Between the Two External Reviewers in the Five-Category RUCAM Score by Occasion

	Occasion 1		Occasion 2	
MAD	Frequency (n = 40)	Percent	Frequency (n = 38)	Percent
0	20	50.0	24	63.2
1	18	45.0	10	26.3
2	2	5.0	3	7.9
3	0	0.0	1	2.6

Table 5. Categorized RUCAM Score for Reviewer A versus Reviewer B Pooled Across the Two Occasions

	Reviewer B					
Reviewer A	Highly Probable	Probable	Possible	Unlikely	Excluded	Totals
Highly probable	5	6	0	0	0	11
Probable	1	31	10	1	0	43
Possible	1	6	6	0	1	14
Unlikely	0	1	4	0	0	5
Excluded	0	1	1	1	2	5
Totals	7	45	21	2	3	78

ever, there was complete agreement in only 58 of 78 reviews (74%). Thus, the ability of the RUCAM to authenticate a DILI event in a clinical research or practice setting is less than ideal.

Discussion

Although this study was conducted by hepatologists with good experience in hepatotoxicity, the RUCAM largely failed to meet the minimum thresholds for a reliable instrument. Discrepancies of three and four points in the continuous version and one point in the categorized version seem rather small; however, reliability is concerned with variability rather than bias. In this regard, there was considerable variability in the RUCAM between the two occasions and among the three reviewers. Under the best-case scenario, the test-retest reliability among the site PIs was only 0.65 while the interrater reliability among the external reviewers was unacceptably low at 0.46. Typically, a test-retest reliability of 0.8 and interrater reliability of 0.6 are expected, and only the U95CL for the former exceeded its threshold.

Multi-item questionnaires are frequently used in other clinical areas to quantify the level of disease activity. These instruments also require clinical judgment, and it is instructive to compare our results against the reliability coefficients of these questionnaires. For example, test-retest reliabilities of 0.85 to 0.93 were reported for the Recent-Onset Arthritis Disability Index[27]; 0.71 to 0.95 for the Dyspnea Management Questionnaire[28]; and 0.87 to 0.97 for the Inflammatory Bowel Disease Questionnaire.[29] Similarly, the interrater reliability of the Pediatric Ulcerative Colitis Activity Index[30] was reported as 0.87, while that of the Myositis Assessment Scale[31] was 0.89. Interestingly, a disease activity index was also developed using a consensus process among clinical experts in idiopathic inflammatory myopathy.[32] Even after an initial training series, however, interrater reliabilities of 0.32 to 0.74 were observed.

There are a number of limitations to the generalizability of these results. First, the ILIAD drugs are well known hepatotoxins and were selected largely for their known DILI signatures. Cases were enrolled only if the site PI felt a priori that there was a significant degree of association between the liver injury and the implicated drug. Moreover, the liver injuries were severe: 73% of cases were hospitalized, many were jaundiced, and 6% required liver transplantation. In effect, these are "classic" DILI cases, and compared to other drugs, should have resulted in greater agreement over time and among the reviewers. On the other hand, the study was conducted retrospectively, with cases going back to 1994. Many medical records and charts for older cases were missing or incomplete, data on death, fulminant hepatic failure, and dechallenge were not always available, and other competing causes may not have been excluded completely. It will be of interest to see if the reliability is greater with more complete data collected prospectively.

Lachin[33] discussed the statistical implications of poor reliability in clinical research. Specifically, the level of association between a measure with poor reliability and other variables as assessed by correlation, regression, or analysis of variance is shrunk toward zero, making it more difficult to declare statistical significance. Statistical power is reduced, so that the sample size must be increased correspondingly. Sensitivity and specificity of the instrument are also attenuated, giving rise to classification errors and impairing its utility to serve as a diagnostic marker. This has significant implications for the RUCAM's ability to detect DILI signals and declare DILI cases.

There are two approaches to overcome these limitations. One is to categorize the instrument. However, our analysis reveals that this maneuver improved matters only marginally. There was complete agreement in only a small majority of cases, and the test-retest and interrater reliabilities remained low. The other is to have m reviewers perform the evaluation independently and take the average. Lachin[33] showed that if ρ is the reliability coefficient of a single assessment, the reliability of the average is given by, $m\rho/[1 + (m - 1)\rho]$. With the three independent reviewers in ILIAD, this would raise the interrater reliabil-

ity from 0.46 to 0.71. This brings the reliability to a more acceptable range and is strongly recommended for research purposes.

Site PIs tended to attribute greater causality score than the external reviewers, which raises important issues. Because the site PI enrolled the case and was the "champion" of that case in the Causality committee, he or she may have been more zealous in attributing the event to a drug-induced liver injury. Alternatively, site PIs may have been more intimately familiar with nuances of the cases not captured completely in the CRF subset and narrative, selectively emphasizing certain components of the instrument. Either way, this suggests that the RUCAM is a "subjective" instrument and casts doubt on its utility as an "objective" measure of DILI causality. It also raises the possibility that causality should only be assessed by reviewers at arm's length from the case. This might avoid bias, but from a reliability perspective, this would be a mistake. The site PIs were consistently more reliable than the external reviewers. Written instructions, criteria for competing causes, and evidence-based revisions, pilot-tested in prospective cohorts, would go a long way toward overcoming these limitations.

Smaller MADs among the three reviewers were observed on the second occasion compared to the first. This may reflect accumulating experience and familiarity with the RUCAM as time progressed. It may also reflect accumulating experience with the "gestalt" of the monthly Causality Committee teleconferences. Nobody wants to be an outlier, and reviewers may have become more adept at anticipating how their colleagues would weigh the evidence and score the case. This weakens the assumption of reviewers working independently as time progressed, and argues that special attention must be paid to this operating assumption.

Finally, there are many who would argue that because of its idiosyncratic nature, the gold standard for adjudicating cases of DILI can only be the clinical judgment of expert hepatologists. Indeed, DILIN is applying an expert opinion process in its clinical studies. However, this is not practical in a clinical setting, and the reliability among practitioners is likely to be lower. Thus, over the long term, priority should be given to developing an authoritative, evidence-based causality instrument that would be easily accessible to the clinical and research communities; for example, over the internet. In the interim, modifications to the RUCAM, including improved instructions, updated criteria for competing causes of liver injury, and a central reference for prior reports of hepatotoxicity, are needed to improve its performance characteristics as an investigational tool.

Acknowledgement: The DILIN expresses its appreciation to Dr. G. Danan for help in developing operating procedures for the RUCAM, and to two reviewers whose comments improved the manuscript considerably.

References

1. Watkins PB, Seeff LB. Drug-induced liver injury: summary of a single topic clinical research conference. HEPATOLOGY 2006;43:618-631.
2. Navarro VJ, Senior JR. Current concepts: drug-related hepatotoxicity. N Engl J Med 2006;354:731-739.
3. Abboud G, Kaplowitz N. Drug-induced liver injury. Drug Saf 2007;30: 277-294.
4. Lee WM, Senior JR. Recognizing drug-induced liver injury: current problems and possible solutions. Toxicol Pathol 2005;33:155-164.
5. Bonkovsky HL, Shedlofsky SI, Jones DP, LaBrecque D. Drug-induced liver injury. In: Boyer TD, Manns MP, Wright TL, eds. Zakim and Boyer's Hepatology—A Textbook of Liver Disease. 5th ed. Philadelphia: Saunders-Elsevier; 2006:503-550.
6. Kaplowitz N. Causality assessment versus guilt-by-association in drug hepatotoxicity. HEPATOLOGY 2001;33:308-310.
7. 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;46:1323-1330.
8. 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;46:1331-1336.
9. Maria VAJ, Victorino RMM. Development and validation of a clinical scale for the diagnosis of drug-induced hepatitis. HEPATOLOGY 1997;26: 664-669.
10. Lucena MI, Camargo R, Andrade RJ, Perez-Sanchez CJ, Sanchez de la Cuesta F. Comparison of two clinical scales for causality assessment in hepatotoxicity. HEPATOLOGY 2001;33:123-130.
11. Stojanovski SD, Casavant MJ, Mousa HM, Baker P, Nahata MC. Atomoxetine-induced hepatitis in a child. Clin Toxicol 2007;45:51-55.
12. Yan B, Leung Y, Urbanski SJ, Myers RP. Rofecoxib-induced hepatotoxicity: a forgotten complication of the coxibs. Can J Gastroenterol 2006;20: 351-355.
13. Fontana RJ, Shakil AO, Greenson JK, Boyd I, Lee WM. Acute liver failure due to amoxicillin and amoxicillin/clavulanate. Dig Dis Sci 2005;50:1785-1790.
14. Cárdenas A, Restrepo JC, Sierra F, Correa G. Acute hepatitis due to Shen-Min: a herbal product derived from *Polygonum multiflorum*. J Clin Gastroenterol 2006;40:629-632.
15. Lynch CR, Folkers ME, Hutson WR. Fulminant hepatic failure associated with the use of black cohosh: a case report. Liver Transpl 2006;12:989-992.
16. Björnsson E, Kalaitzakis E, Klinteberg VAV, Alem N, Olsson R. Long-term follow-up of patients with mild to moderate drug-induced liver injury. Aliment Pharmacol Ther 2007;26:79-85.
17. Andrade RJ, Lucena MI, Kaplowitz N, García-Muṇoz B, Borraz Y, Pachkoria K, et al. Outcome of acute idiosyncratic drug-induced liver injury: long-term follow-up in a hepatotoxicity registry. HEPATOLOGY 2006;44: 1581-1588.
18. Masumoto T, Horiike N, Abe M, Kumaki T, Matsubara H, Fazle Akbar SM, et al. Diagnosis of drug-induced liver injury in Japanese patients by criteria of Consensus Meetings in Europe. Hepatol Res 2003;25:1-7.
19. Lee WM, Larrey D, Olsson R, Lewis JH, Keisu M, Auclert L, et al. Hepatic findings in long-term clinical trials of ximelagatran. Drug Saf 2005;28: 351-370.
20. Andrade R, Lucena MI, Alonso A, García-Cor&tacute;es M, García-Ruiz E, Benitez R, et al. HLA class II genotype influences the type of liver injury in drug-induced idiosyncratic liver disease. HEPATOLOGY 2004;39:1603-1612.

21. Hoofnagle JH. Drug-Induced Liver Network (DILIN). HEPATOLOGY 2004;40:773.
22. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;327:307-310.
23. Eliasziw M, Young SL, Woodbury MG, Fryday-Field K. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. Phys Ther 1994;74: 777-788.
24. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat Med 2000;19:1141-1164.
25. Walsh JF, Reznikoff M. Bootstrapping: a tool for clinical research. J Clin Psychol 1990;46:928-930.
26. Donner A, Eliasziw M. Sample size requirements for reliability studies. Stat Med 1987;6:441-448.
27. Salaffi F, Stancati A, Neri R, Grassi W, Bombardieri S. Measuring functional disability in early rheumatoid arthritis: the validity, reliability and responsiveness of the Recent-Onset Arthritis Disability (ROAD) index. Clin Exp Rheumatol 2005;23(Suppl 39):S31-S42.
28. Migliore Norweg A, Whiteson J, Demetis S, Rey M. A new functional status outcome measure of dyspnea and anxiety for adults with lung disease: the dyspnea management questionnaire. J Cardiopulm Rehabil 2006; 26:395-404.
29. Cheung WY, Garratt AM, Russell IT, Williams JG. The UK IBDQ—a British version of the inflammatory bowel disease questionnaire: development and validation. J Clin Epidemiol 2000;53:297-306.
30. Turner D, Otley AR, Mack D, Hyams J, de Bruijne J, Uusoue K, et al. Development, validation, and evaluation of a pediatric ulcerative colitis activity index: a prospective multicenter study. Gastroenterology 2007; 133:423-432.
31. Huber AM, Feldman BM, Rennebohm RM, Hicks JE, Lindsley CB, Perez MD, et al. Validation and clinical significance of the Childhood Myositis Assessment Scale for assessment of muscle function in the juvenile idiopathic inflammatory myopathies. Arthritis Rheum 2004;50:1595-1603.
32. Isenberg DA, Allen E, Farewell V, Ehrenstein MR, Hanna MG, Lundberg IE, et al. International consensus outcome measures for patients with idiopathic inflammatory myopathies. Development and initial validation of myositis activity and damage indices in patients with adult onset disease. Rheumatology 2004;43:49-54.
33. Lachin JM. The role of measurement reliability in clinical trials. Clin Trials 2004;1:553-566.