UCSF

UC San Francisco Previously Published Works

Title

Approaches to Research Determination of Late Acute Cellular Rejection in Pediatric Liver Transplant Recipients

Permalink

https://escholarship.org/uc/item/5d96d87c

Journal

Liver Transplantation, 27(1)

ISSN

1527-6465

Authors

Mazariegos, George V
Shneider, Benjamin L
Shemesh, Eyal
et al.

Publication Date

2021

DOI

10.1002/lt.25903

Peer reviewed

Approaches to research determination of late acute cellular rejection in pediatric liver transplant recipients

George V. Mazariegos[1], **Benjamin L. Shneider**[2], **Eyal Shemesh**[3], **Deborah Schady**[4], **Hector Melin-Aldana**[5], **Soo-Jin Cho**[6], **Ravinder Anand**[7], **Jinson Erinjeri**[7], **Rachel Annunziato**[3], **Miguel Reyes-Mugica**[8]

[1]Hillman Center for Pediatric Transplantation, Thomas E. Starzl Transplantation Institute, UPMC Children's Hospital of Pittsburgh, PA.

[2]Department of Pediatrics, Baylor College of Medicine, Houston, TX

[3]Icahn School of Medicine at Mount Sinai and the Department of Pediatrics, Mount Sinai Health System, New York, NY

[4]Department of Pathology, Baylor College of Medicine, Houston, TX

[5]Department of Pathology and Laboratory Medicine, Northwestern University Feinberg School of Medicine and Ann and Robert H. Lurie Children's Hospital, Chicago, IL

[6]Department of Pathology, UCSF, San Francisco, CA.

[7]The Emmes Company, LLP; Rockville, MD, USA.

[8]Hillman Center for Pediatric Transplantation Department of Pathology, UPMC Children's Hospital of Pittsburgh, Pittsburgh, PA.

Abstract

A "central pathology" or "site" reading of biopsy slides is used in liver transplant clinical trials to determine rejection. We evaluated inter-rater reliability of readings of "rejection or not" using digitized slides from the Medication Adherence in Pediatric Liver Transplant Recipients (MALT) study. 4 masked experienced pathologists read the digitized slides and then reread them after a study-specific histologic end-point development program. Agreement was expressed throughout as a Kappa or Fleiss Kappa statistic (κ). A $\kappa > 0.6$ was predefined as desirable. Readings were correlated with immunosuppressant adherence (The Medication Level Variability Index, MLVI), and maximal liver enzyme levels during the study period. Interrater agreement between site and central review in MALT, and between 4 pathologists later on, was low ($\kappa=0.44$, Fleiss $\kappa = 0.41$, respectively). Following the end-point development program, agreement improved and became acceptable ($\kappa = 0.71$). The final reading was better-aligned with maximal GGT levels and MLVI

Corresponding author: George V Mazariegos, M.D. George.Mazariegos@chp.edu, Chief, Pediatric Transplantation, Professor of Surgery and Critical Care, University of Pittsburgh School of Medicine, Thomas E. Starzl Transplantation Institute and Hillman Center for Pediatric Transplantation, UPMC Children's Hospital of Pittsburgh, 4401 Penn Ave., Pittsburgh, PA 15224 United States, Phone: (412) 692-6110, Fax: (412) 692-6117.

as compared with the original "central" reading. We found substantial disagreement between experienced pathologists reading the same slides. A unique study-specific procedure improved interrater reliability to the point it was acceptable. Such a procedure may be indicated to increase reliability of histopathologic determinations in future research, and perhaps also clinically.

Keywords

Transplantation; Pathology; Rejection; Adherence; Histology

INTRODUCTION

Prevention and management of late acute cellular rejection (LAR)[1] has assumed greater clinical importance in pediatric liver transplantation, as LAR has been recognized as detrimental to long term outcomes.[2–4] Not surprisingly therefore, "rejection" is frequently used as the primary outcome of interest in clinical trials as well as in practice.[5–7] But it is not entirely clear how to determine whether a patient has rejection or not. Histology results are the current standard. However, differences in the way pathologists read the same specimen are often not considered. To enhance reproducibility in clinical trials, histological findings are sometimes determined by a "central pathology" procedure in which the same pathologist (masked to clinical information) reviews all study histology. Alternatively, site-determined histological findings are used as the standard, but precisely what information, in particular nonhistological information, sites use to arrive at the definition of rejection is rarely clearly defined. To illustrate this point further, we conducted a clinicaltrials.gov search, in which we identified at all completed registered trials in pediatric liver transplantation in which rejection was either a primary (8 trials) or secondary (10 trials) outcome.[8] Those trials used only one pathologists' reading without an attempt to verify that this reading was reproducible. Three used a single "central" reading, and the rest used site readings.

Currently, therefore, in pediatric liver transplantation, the field lacks validation of a robust, reproducible outcome measure to gauge the efficacy of interventions aiming to reduce the incidence of LAR, because only one pathologist's reading of a histologic specimen is accepted as an outcome measure, but the interrater reliability of pathologists' determination of LAR is not known. If interrater reliability is poor (different pathologists can reach different conclusions when reading the same slide), then neither a "central" pathologist nor a "local" reading would be appropriate for use in trials. In addition, the most frequently used outcome, a local determination of rejection, is potentially fraught with inconsistencies: there is no standard process to determine what information should be used to supplement the histopathologic reading, sites' clinical practices vary, and the site determination is likely to be biased by the patient's known behavior and clinical course.

The field, therefore, is in urgent need for a thorough investigation of the degree of interrater reliability in pathologists' determinations of "rejection". We must develop and validate a robust, objective, unbiased process that could reliably and consistently inform us of the effect of a given intervention on the prevalence of rejection. So long as we worry that

different readers might assign significantly different determinations to the same exact biopsy slide, and thus completely change study results, the field cannot reasonably move forward in developing and testing new interventions.

The MALT (Medication Adherence in Children who had a Liver Transplant; ClinicalTrials.Gov registration # NCT01154075) study was a prospective cohort study that enrolled pediatric liver transplant recipients in 5 centers across the United States, aiming to evaluate the relationship between nonadherence to tacrolimus and LAR.[9] In the design of MALT, it was appreciated that the clinical determination of rejection could have been biased by the readers' knowledge of patient behavior (e.g. nonadherence). As a result, central assignment of late acute cellular rejection, done by a pathologist who was masked to the clinical course of the patient, was used in MALT. But after MALT was concluded, it was found that there was poor correlation between site and central assignment of rejection status. As the investigation moved towards the design of an intervention study (iMALT, NCT03691220) that aims to improve adherence, it was essential to develop a robust outcome measure that was unbiased by clinical information on one hand, but also displayed acceptable inter-rater reliability. The present work was done at the iMALT study design phase to create a reliable mechanism for histological determination of rejection, using available masked slides from the previous MALT cohort.

In the present investigation a group of 4 expert pediatric liver histo-pathologists read scanned digital slides of the liver biopsies obtained in MALT. Their reading was masked to any clinical information. In light of the poor inter-rater reliability of these initial readings, we created a study-specific histologic end-point development program with a goal of enhancing the inter-rater reliability for the study determination of LAR. A year later, the same set of slides was re-read by the same 4 pathologists. Those first and second phase readings were also compared with prior site and central readings of rejection in the original study (MALT), as well as with clinical information which was not known to the pathologists. This information included maximal levels of serum GGT and ALT and the level of fluctuation of tacrolimus blood levels over time (the Medication Level Variability Index [MLVI], a measure of adherence to medications)[9] - parameters which were previously chosen as other end-points in MALT.

The primary aim of the study was to investigate interrater reliability between pathologists in the determination of "rejection", and evaluate whether a study-specific histologic end-point development program can increase this agreement in a future masked reading by the same pathologists. A secondary aim was to evaluate the correlation between clinical parameters (MLVI and liver enzymes) and pathologists' readings, as a way to gauge which of those readings were better correlated with other markers that are known to be associated with rejection.

METHODS

Original MALT cohort

The MALT (Medication Adherence in pediatric Liver Transplant recipients; ClinicalTrials.Gov registration # NCT01154075) study was a prospective cohort study that

enrolled pediatric liver transplant recipients in 5 centers across the United States. 400 participants were followed for two years, to determine the relationship between nonadherence to tacrolimus and LAR.[9] When any participant in MALT had a clinically indicated biopsy (research and/or protocol biopsies were not performed), one hematoxylin and eosin (H&E) stained slide, chosen by the site as representative of the case, was read by a central pathology team in the absence of clinical information. The study-specific histological end-point adopted in MALT and all later readings throughout the present study was the presence or absence of rejection based on features listed by the International Working Party on Terminology for Hepatic Allograft Rejection, published as a 2016 Comprehensive Update of the Banff Working Group on Liver Allograft Pathology.[10] These criteria include: (1) mixed but predominantly mononuclear portal inflammation, containing "blastic" (activated) lymphocytes, neutrophils, and frequently eosinophils; (2) bile duct inflammation/ damage; (3) subendothelial inflammation of portal or terminal hepatic veins. We did not use the scoring mechanism published alongside those recommendations, because we did not aim to determine rejection severity.

Centrally-read biopsy (rejection or not), the site determinations of acute rejection (based on biopsy results as read at the site in the context of clinical information available as part of routine clinical care), serum ALT / GGT, and trough tacrolimus levels, were all recorded in MALT for a period of 2 years for each of the participants. Adherence to tacrolimus in MALT was determined by the Medication Level Variability Index (MLVI), a calculation of the standard deviation of tacrolimus blood levels over time.[9] A higher MLVI means a higher degree of fluctuation between individual blood levels in a given patient, which translates into more erratic adherence to medications. A high MLVI (low adherence) predicts poor post-transplant outcomes.[9,11]

The original MALT readings were based on glass slides. For subsequent readings (after MALT), slides were scanned using Aperio CS2 or AT2 scanners, Leica Biosystems® (Department of Pathology, UPMC Children's Hospital of Pittsburgh).

Sample size

The primary analysis in MALT, by design, only looked at MLVI prior to the first episode of rejection and correlated with the presence or absence of centrally assigned rejection. Seventy-four biopsies were performed and 74 "cases" read by both the central and site pathologists during the 2 year follow up period of the 400 participants in the original MALT cohort; some participants had more than one biopsy. The results of the central and site review of the slides were distinct and not shared. Nine slides were not available for scanning, and thus 65 scanned slides from biopsies performed on 50 participants were available for the subsequent readings (Table 1, Figure 1).

Procedures

The degree of agreement between the original masked MALT central pathology reading and the local unmasked reading, which was done as a part of clinical care at the site ("central" versus "site" reading), was assessed using the Kappa statistic (κ), applied to the 74 biopsies which were reviewed in MALT. The maximal ALT / GGT level recorded during study

was compared amongst participants who had site vs. central assignment of rejection, and participants who had no rejection but underwent a clinically indicated liver biopsy.

Two years after the conclusion of the MALT study in preparation for the interventional study iMALT, a second masked reading of the aforementioned 65 digitally scanned biopsy slides from MALT, was undertaken by four experienced pediatric liver pathologists at 4 different sites (**iMALT1**). Pathologists were asked to determine whether they observed potentially actionable (i.e. for which a therapeutic intervention would be considered by the team caring for the patient) late acute cellular rejection. The degree of agreement between those four readings was assessed including the number of times in which a full (all readings are the same) or partial (majority of readers reach the same conclusion) agreement was reached. The low level of agreement in **iMALT1** led to the development of a study-specific histologic end-point development program,for the purpose of reaching a consistent way of determining the presence or absence of rejection in the upcoming iMALT clinical trial.

iMALT LAR study-specific histologic end-point development program

The 4 experienced pediatric liver pathologists participated in this study-specific program which was led by an independent senior pathologist (Anthony J Demetris, M.D.). It included a general discussion of the features of rejection that would meet the criteria of LAR in iMALT. This included a brief review of Banff features and their relevance to the iMALT study aims. A key part of the program was a collaborative review and discussion on-line of *three* slides, chosen at random, in which there were discrepant readings in **iMALT1**. The collaborative review, included a standard and detailed review of each and every one of the different elements of rejection as identified in the Banff consensus statement[10] (but without explicit scoring of the elements): bile duct epithelial damage; portal endothelialitis; eosinophilic portal infiltration; and venular endothelialitis. The discussion centered not only on whether or not those elements were present, but also on reaching a consensus about how much of those need to be present in order to make a determination of "rejection" for the purpose of the particular proposed research in iMALT.

A year later, the same digitally scanned slides were re-read by the same four pathologists,with the goal of applying the methods discussed in the study-specific histologic end-point development program to the dichotomous determination of rejection or no rejection to the same set of 65 digitized slides (**iMALT2**). This reading was again done independently but at the same time as part of a video conference amongst the four pathologists who were not aware of each other's readings.

iMALT 2 adjudication readings

Ultimately, a final assignment of rejection or not was made for all 65 slides (**iMALT2 adjudication**). This included an open on-line discussion of slides (n = 6) where there was an even split on the finding (e.g. 2 in favor and 2 against rejection). The endpoint of the adjudication process was a final determination of "rejection" or "not rejection" for all 65 slides (**iMALT2 adjudication** results).

Participant outcomes

To evaluate the degree to which the above readings were associated with patient outcomes, we evaluated the following outcomes experienced by the 50 participants who underwent clinically indicated biopsies. For those participants who had more than one biopsy performed during the study, if any biopsy was read as "rejection", they were assigned into the "rejection" group:

1. MLVI (adherence) score, using data from the entire 2 year follow-up.

2. Maximal gamma glutamyl transferase (GGT) level during the two years in MALT.

3. Maximal alanine aminotransferase (ALT) level during the two years in MALT.

Statistical Analyses

Descriptive analyses include the number of biopsies read, the number of cases in which a "rejection" was assigned by various reading stages, and the number and percent of cases in which an assignment of "rejection" was associated with increased MLVI (nonadherence) and with the clinical outcomes and treatment outcomes as presented above.

To compute the significance of the difference between rejection versus "no rejection" assignments and MLVI outcomes, we used a non-parametric Kruskal-Wallis test when MLVI is treated as a continuous variable (higher MLVI = worse adherence) or a Fisher's Exact Test when MLVI is treated as a threshold variable (in which MLVI values equal to or greater than 2.5 are considered to be "nonadherence", as previously described). A non-parametric Kruskal-Wallis test was used for computing significance between maximal ALT/GGT.

A Kappa statistic (κ) was used to determine the degree of agreement between individual raters. A Fleiss Kappa was calculated to denote the degree of agreement between multiple raters. The differences between kappas were compared using a two-sided t-test. A $\kappa>0.6$ was predefined as desirable. Kappa statistic (Cohen's) was determined using SAS (Version 9.4) software. Fleiss Kappa was determined using R (version 3.6.1) software with IRR package.

RESULTS

Central vs Site - read Rejection in MALT

A total of 74 biopsies performed during the conduct of MALT were reviewed during MALT by the sites and by the central pathologist. For 53 of the 74 biopsies performed for clinical indications during MALT, the central pathologist assigned rejection, while the site assigned 44 as rejection (Table 2). Substantial discrepancies were observed between the central and local assignment of rejection for these biopsies. There was concordance of rejection for 39 biopsies and no rejection for 16, leaving discordant readings in the remaining 19 biopsies. The Kappa coefficient for this comparison was 0.44.

Four Pathologists' reading of Biopsies (iMALT), before and after the study-specific histologic end-point development program

To shed further light on the histopathologic findings, a group of 4 pathologists including the central pathologist in MALT reread the 65 available digitized biopsy slides (iMALT readings). Raw results are presented in Supplemental Table 1. In the first review of these biopsies by the four pathologists (iMALT1), substantial discordance was observed in the assignment of rejection (Fleiss kappa = 0.41). There was universal concordance of rejection and absence of rejection in 20 and 12 biopsies, respectively. In 10 cases, 3 readers of the 4 assigned rejection; in 10 cases, 3 of 4 readers assigned no rejection, while in 14 cases, assignments were evenly split relative to rejection. After the training, concordance in readings (iMALT2) improved (Fleiss kappa = 0.71). Individual Kappas for the first and second readings are presented in Figure 2. Intra-rater reliability between the first and second readings were κ= 0.81, 0.56, 0.52 and 0.79 for each of the 4 pathologists. The group's assignment status changed for 23 of the 65 biopsies. There was universal concordance of rejection in 28 cases in the first reading, which improved to 36 cases in the second reading and similarly, complete agreement on the absence of rejection was observed in 20 cases in the first reading and 28 cases in the second reading (Supplemental Table 1).

Remaining discrepant readings (split decision) during the second reading were then discussed by teleconference with an on-line review of the relevant digitized slides (n=6) leading to the final adjudicated assignments of rejection from these biopsies (**iMALT2 adjudication**). The agreement between the resulting iMALT2 adjudication reading and the original MALT central pathology reading (κ= 0.47), or the site reading (κ= 0.49) was weak. Tables 3 and 4 describe the "adjudication" reading (iMALT2 adjudication) as compared with the central and site readings in MALT to determine relationship to clinical parameters of interest (MLVI and maximal ALT / GGT). MLVI, either as a continuous variable or relative to a relevant cut-off, was significantly higher in site and iMALT2 adjudication assignment of rejection versus no rejection. Maximal ALT was only higher in rejection versus no rejection for the site assignments, while maximal GGT was only higher in rejection versus no rejection for the iMALT2 adjudication assignment.

DISCUSSION

We assessed different methods for the histologic assignment of "rejection" in the MALT study cohort, and found that the masked central pathology assessment was not well-aligned with site determinations. We also found that a separate review of slides by 4 expert pediatric liver pathologists initially had a high level of disagreement between the readers. However, the inter-rater agreement substantially increased after a study-specific histologic end-point development program that involved open discussions about a few slides and a review of the aims of the intervention trial (iMALT) for which the training was designed. The goal in iMALT is to identify rejection as a dichotomous unscored outcome. The study seeks to identify histologic findings, which would in the absence of other information prompt clinical intervention. When looking at clinical correlates of rejection, we found that the last reading (iMALT2) showed not only a better interrater agreement but also, after adjudication of the few "split" readings (iMALT2 adjudication) better alignment with clinical information

compared to the central assignment of rejection (Tables 3 and 4). Our results suggest that study-specific histologic end-point development programs could be one way to mitigate disagreement between raters, as well as to ensure a good alignment between histological readings.

The substantial discrepancy in the assignment of rejection status that we observed in the initial readings is worrisome given both the clinical and research importance of the diagnosis of rejection. A highly relevant question is which reading is "correct," which also begs the question as to the definition of rejection. Rejection is an immunologic process, which potentially leads to liver injury that may necessitate enhanced immunosuppression. Is rejection simply a histologic determination or is it a clinico-pathologic entity? Should that clinico-pathologic entity incorporate clinical information prior to and potentially after a clinical intervention is made? In the absence of a true "gold standard" for late acute cellular rejection, we anchored our findings to biochemical evidence of liver injury in the form of maximal ALT and GGT, presuming these would be increased in rejection. Others, for example, linked interface hepatitis and perivenular inflammation (cluster 1 patients) with a 13 gene signature previously associated with T cell mediated rejection as well as specific genes linked with TCMR in clinical and experimental settings (eg, CXCL9 , CXCL10)1.[12] Examination of markers such as these could add mechanistic and diagnostic insight in the future if they are further studied and validated.

The study-specific histologic end-point development program yielded enhanced agreement and the important finding of a significant difference (2-fold increase) in GGT between those with and without rejection. This suggests that the training resulted in a more consistent reading that was also better aligned with the process that we consider to be late acute cellular rejection.[13,14]

Although ours is perhaps the largest study in pediatric liver transplantation to note such differences, substantial disagreements between histological readings have been previously reported, both in the particular setting of liver biopsies[15–17] and more generally with other pathological findings.[18,19]

The most important insight from our findings is that in a research setting, the use of biopsy readings as the sole arbiter of the presence or absence of LAR is probably rarely appropriate. A "central pathologist" reading is likely to result in irreproducible findings, unless a study-specific histologic end-point development program is added before the study commences. Nevertheless, single-pathologist readings of a biopsy – whether as a "site" reading or a "central" reading – is commonly and perhaps exclusively used[8] to determine treatment effects. We have identified a way to improve the reliability of such readings via a dedicated program that takes into account the specific study design and needs. Presumably, the use of such procedures will also include a re-evaluation of readers' consistency as we have done. What should be the best way to evaluate interrater variability in such readings and what level of agreement should be deemed "satisfactory"? We are not sure that there is a single answer to this question, and it is possible that each study, and each clinical setting, should develop its own "minimal standard" for reproducibility. The original interpretation of Kappa posits that Kappas between 0.21–0.4 denote "fair agreement", 0.41–0.60 is "moderate agreement",

0.61–0.8 is "substantial," and above 0.81 is "almost perfect".[20] A Kappa of 0.6 is generally considered a requisite minimum for clinical decision-making, but there is little empirical research to support this assertion, and it is possible that a higher threshold should be required, especially in cases in which a specific determination might lead to treatment modifications that may result in adverse consequences if done wrong (as in our case). Our results, however, were not subtle: our "pre-consensus" Kappas, in the range of 0.4–0.5, are most certainly unacceptable in our setting, as demonstrated by a case-by-case review (Supplemental Table 1) that showed a large number of cases in which there was disagreement. Whether or not the final Kappa of 0.7 is good enough could be debated.

Our results have several limitations. First, the iMALT readings used scanned slides while MALT readings used actual slides. Scanned slides are commonly used in both clinical practice and research. Published information in transplant medicine, using the same equipment that we used, suggests that histologic reviews of scanned slides are at least as good in detecting rejection as reviews of glass slides.[21] We found that it was possible to increase interrater agreement while using the same scanned slides, and that very substantial disagreement between readers (of the same magnitude) was observed when either glass slides or scanned slides were used. The existing literature combined with our own findings, therefore, strongly suggest that using scanned versus actual slides was not a major explanatory factor in the present study. Second, the readings at the central and iMALT procedures were based on a chosen single slide, while site readings may have involved one or more slides. Third, our results pertain to children, and it is not entirely clear whether our results are also applicable to adults, although they may be.

The fact that the present sample was obtained from a nationally representative multisite cohort[22] and involved readings by leading pathologists, all of whom have specific expertise in the study population, suggests that in spite of the limitations mentioned above, our results are likely to be generalizable.

Clinical trials need an unbiased, masked outcome measure to determine intervention effects, and site readings can be unacceptably biased in this regard. It is reassuring, therefore, that a study-specific histologic end-point development program could substantially increase both reliability as well as clinical relevance of the readings, even while still masked. This process should be strongly considered in future clinical trials that rely on histologic features of liver rejection as a critical endpoint.

The diagnosis of liver allograft rejection in the clinical setting does include ancillary information in addition to histologic features but is still typically assessed by a single observer. A clinician can evaluate the patient's course after the clinical determination of rejection status was made – changing course given new information related to treatment responses. In this way, the uncertainty that we report related to the reading of a single pathologist can be mitigated, but perhaps not entirely eliminated, in clinical practice.

Our novel study-specific histologic end-point development program methodology provides a way to mitigate the problem of inter-rater disagreement in the context of (necessary) masking. The concepts of interrater reliability and pre-trial end-point development processes

to improve such reliability, which are reviewed in this manuscript, are relevant to studies of patients with low immunosuppression (whether due to nonadherence or other reasons), but also more broadly. We believe that our results are also important in developing protocols to better study and define rejection, either as an endpoint or an outcome measure, including that of the subclinical type.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS/FUNDING

DATA AVAILABILITY STATEMENT

The authors confirm that all relevant data are included in the article and/or its supplementary information files.

List of abbreviations:

|	
UPMC	University of Pittsburgh Medical Center
MALT	Medication Adherence in Pediatric Liver Transplant Recipients study
MLVI	Medication Level Variability Index
ҡ	Kappa or Fleiss Kappa Statistic
GGT	Gamma Glutamyl Transferase
LAR	Late Acute Cellular Rejection
ALT	Alanine Aminotransferase
H&E	Hematoxylin / Eosin

References

1. Nacif LS, Pinheiro RS, Pecora RA, et al. Late acute rejection in liver transplant: a systematic review. Arq Bras Cir Dig. 2015;28(3):212–215. [PubMed: 26537150]

2. Anand AC, Hubscher SG, Gunson BK, McMaster P, Neuberger JM. Timing, significance, and prognosis of late acute liver allograft rejection. Transplantation. 1995;60(10):1098–1103. [PubMed: 7482715]

3. Uemura T, Ikegami T, Sanchez EQ, et al. Late acute rejection after liver transplantation impacts patient survival. Clin Transplant. 2008;22(3):316–323. [PubMed: 18190550]

4. Thurairajah PH, Carbone M, Bridgestock H, et al. Late acute liver allograft rejection; a study of its natural history and graft survival in the current era. Transplantation. 2013;95(7):955–959. [PubMed: 23442806]

5. Florman S, Schiano T, Kim L, et al. The incidence and significance of late acute cellular rejection (>1000 days) after liver transplantation. Clin Transplant. 2004;18(2):152–155. [PubMed: 15016129]

6. D'Antiga L, Dhawan A, Portmann B, et al. Late cellular rejection in paediatric liver transplantation: aetiology and outcome. Transplantation. 2002;73(1):80–84. [PubMed: 11792983]

7. Sundaram SS, Melin-Aldana H, Neighbors K, Alonso EM. Histologic characteristics of late cellular rejection, significance of centrilobular injury, and long-term outcome in pediatric liver transplant recipients. Liver Transpl. 2006;12(1):58–64. [PubMed: 16382471]

8. ClinicalTrials.gov. Study identifiers: NCT-00149810, 01207141, 01154075, 00195988, 01428700, 01960322, 00367146, 00320606, 01638559, 00282256, 00927992, 01598987, 00531921, 01371331, 01137084, 02124382, 00395018, 00074386. In:List.

9. Shemesh E, Bucuvalas JC, Anand R, et al. The Medication Level Variability Index (MLVI) Predicts Poor Liver Transplant Outcomes: A Prospective Multi-Site Study. Am J Transplant. 2017;17(10):2668–2678. [PubMed: 28321975]

10. Demetris AJ, Bellamy C, Hubscher SG, et al. 2016 Comprehensive Update of the Banff Working Group on Liver Allograft Pathology: Introduction of Antibody-Mediated Rejection. Am J Transplant. 2016;16(10):2816–2835. [PubMed: 27273869]

11. Lieber SR, Helcer J, Shemesh E. Monitoring drug adherence. Transplant Rev (Orlando). 2015;29(2):73–77. [PubMed: 25432088]

12. Feng S, Bucuvalas JC, Demetris AJ, et al. Evidence of Chronic Allograft Injury in Liver Biopsies From Long-term Pediatric Recipients of Liver Transplants. Gastroenterology. 2018;155(6):1838–1851 e1837. [PubMed: 30144432]

13. Pappo O, Ramos H, Starzl TE, Fung JJ, Demetris AJ. Structural integrity and identification of causes of liver allograft dysfunction occurring more than 5 years after transplantation. Am J Surg Pathol. 1995;19(2):192–206. [PubMed: 7832279]

14. Portmann B, Slapak GI, Gane E, Williams R. Pathology and biopsy diagnosis of the transplanted liver. Verh Dtsch Ges Pathol. 1995;79:277–290. [PubMed: 8600691]

15. Rousselet MC, Michalak S, Dupre F, et al. Sources of variability in histological scoring of chronic viral hepatitis. Hepatology. 2005;41(2):257–264. [PubMed: 15660389]

16. Russo P, Magee JC, Boitnott J, et al. Design and validation of the biliary atresia research consortium histologic assessment system for cholestasis in infancy. Clin Gastroenterol Hepatol. 2011;9(4):357–362 e352. [PubMed: 21238606]

17. Demetris AJ, Seaberg EC, Batts KP, et al. Chronic liver allograft rejection: a National Institute of Diabetes and Digestive and Kidney Diseases interinstitutional study analyzing the reliability of current criteria and proposal of an expanded definition. National Institute of Diabetes and Digestive and Kidney Diseases Liver Transplantation Database. Am J Surg Pathol. 1998;22(1):28–39. [PubMed: 9422313]

18. Vennalaganti P, Kanakadandi V, Goldblum JR, et al. Discordance Among Pathologists in the United States and Europe in Diagnosis of Low-Grade Dysplasia for Patients With Barrett's Esophagus. Gastroenterology. 2017;152(3):564–570.e564. [PubMed: 27818167]

19. Adamson AS, Welch HG. Machine Learning and the Cancer-Diagnosis Problem - No Gold Standard. N Engl J Med. 2019;381(24):2285–2287. [PubMed: 31826337]

20. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276–282. [PubMed: 23092060]

21. Ozluk Y, Blanco PL, Mengel M, Solez K, Halloran PF, Sis B. Superiority of virtual microscopy versus light microscopy in transplantation pathology. Clin Transplant. 2012;26(2):336–344. [PubMed: 21955102]

22. Shemesh E, Mitchell J, Neighbors K, et al. Recruiting a representative sample in adherence research-The MALT multisite prospective cohort study experience. Pediatr Transplant. 2017;21(8).
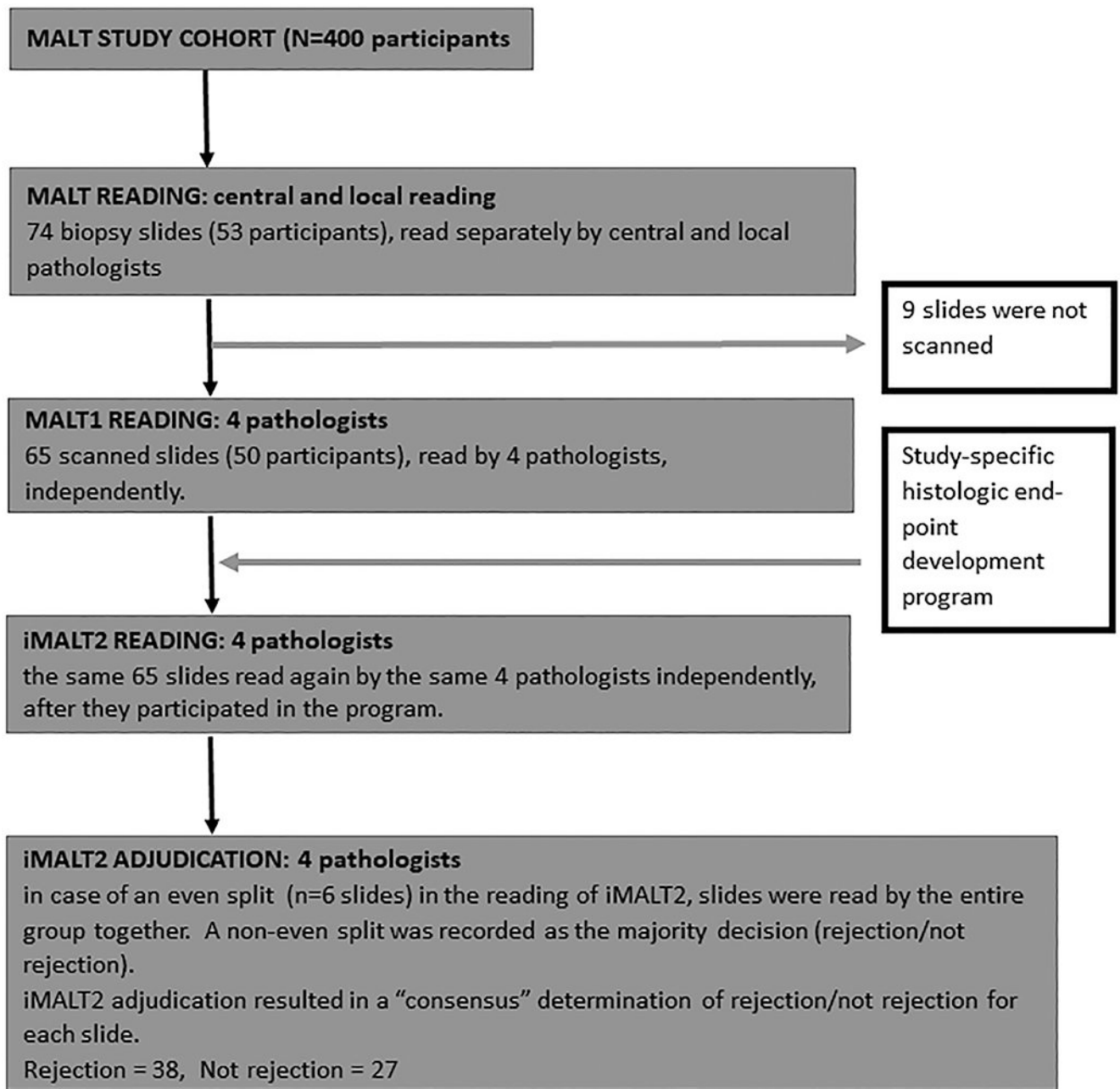
Figure 1.
Diagram of the flow of assignment of rejection for the biopsy slides analyzed.
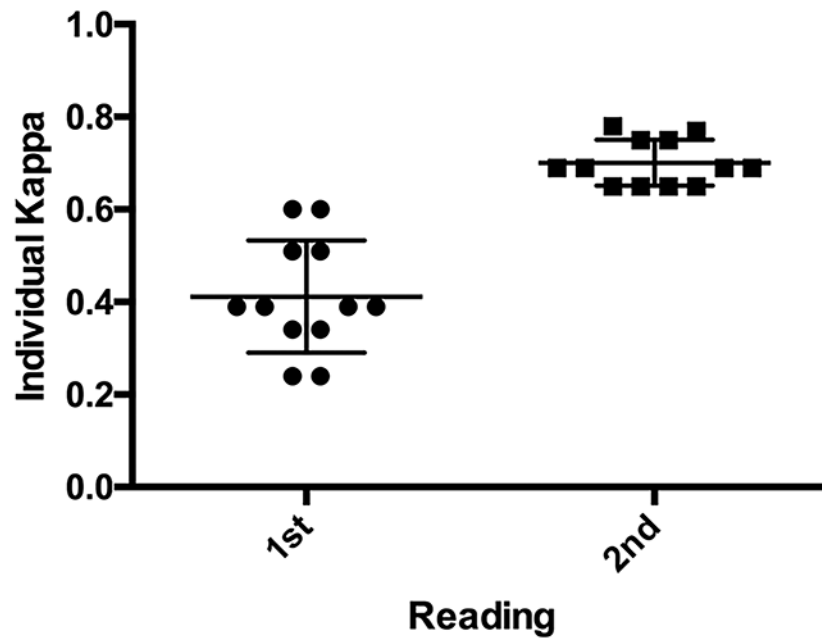
Figure 2.
Kappas for the iMALT1 and iMALT2 readings. Bars represent mean (Fleiss) Kappa, ± Standard Deviation. The differences between the two reading Kappas are significant, p<0.0001.

TABLE 1:

DIFFERENT READINGS IN THE CURRENT ANALYSIS:

READING DESIGNATION	PROCEDURE	TIMING (YEAR)	# OF SLIDES	PURPOSE
MALT "central"	The original masked reading by MALT central pathology	During the MALT cohort study: 2009-2015	74	Central assignment of "rejection", without clinical information
MALT "site"	The original clinical reading by the site at the time the biopsy was done	During the MALT cohort study: 2009-2015	74	Routine clinicopathologic diagnosis of rejection.
*Second reading (**iMALT1**)*	4 experienced pathologists, individual masked reading of available scanned MALT slides	2017	65	Evaluate degree of agreement in a masked reading
Study Specific Histologic End-point Development Program				
*Third reading (**iMALT2**)*	The same 4 experienced pathologists, individual masked reading of MALT slides, after program	2018	65	Evaluate degree of agreement in a masked reading after consensus building
*Third reading "adjudication" - on a rolling basis, immediately following the individual reading above (**iMALT2 adjudication**)*	The same 4 experienced pathologists, open reading and discussion of MALT slides in cases of an even "split" rejection vs no rejection (n=6)	2018	65	Each slide was determined to be "rejection" vs "non rejection" either by a majority read or adjudication when needed. This final readiing was compared with previous MALT readings.

TABLE 2

Relationship and Agreement between Central Pathologist and Site Biopsy Assigned Rejection

| | Central Pathologist Assigned Rejection | | | |
	Yes	No	Total	Kappa Coefficient (95% C.I.)
Site Biopsy Assigned Rejection				
No	14	16	30	0.44 (0.23-0.65)
Yes	39	5	44	
Total	53	21	74	

TABLE 3:

Correlation of Adherence as Measured by the Medication Level Variability Index (MVLI) Relative to Different Assignments of Rejection

MLVI	Site Assignment of Rejection			Central Pathology Assignment of Rejection			iMALT2 Adjudication Assignment of Rejection		
	No N= 17	Yes N = 33	p-value	No N = 11	Yes N =39	p-value	No N = 19	Yes N = 31	p-value
MLVI (Continuous Variable)									
Mean (SD)	2.2 (1.2)	3.3 (1.7)	0.006	2.5 (1.3)	3.0 (1.7)	0.23	2.3 (1.3)	3.2 (1.7)	0.02
Median (Min, Max)	2.3 (0.7, 5.6)	3.1 (0.5, 9.5)		2.4 (0.7, 5.6)	2.9 (0.5, 9.5)		2.3 (0.5, 5.9)	3.1 (0.9, 9.5)	
MLVI 2.0									
Yes	9 (52.9%)	28 (84.8%)	0.01	7 (63.6%)	30 (76.9%)	0.37	11 (57.9%)	26 (83.9%)	0.04
No	8 (47.1%)	5 (15.2%)		4 (36.4%)	9 (23.1%)		8 (42.1%)	5 (16.1%)	
MLVI 2.5									
Yes	5 (29.4%)	24 (72.7%)	0.003	5 (45.5%)	24 (61.5%)	0.34	6 (31.6%)	23 (74.2%)	0.003
No	12 (70.6%)	9 (27.3%)		6 (54.5%)	15 (38.5%)		13 (68.4%)	8 (25.8%)	

TABLE 4:

Correlation of Maximal Gamma Glutamyl Transferase and Alanine Amino Transferase Levels Relative to Different Assignments of Rejection

Liver Enzyme	Site Assignment of Rejection			Central Pathology Assignment of Rejection			iMALT2 Adjudication Assignment of Rejection		
	No N=17	Yes N=33	p-value	No N=11	Yes N=39	p-value	No N=19	Yes N=31	p-value
Maximum GGT (U/L)									
Mean (SD)	263.9 (216.2)	427.3 (458.2)	0.25	213.3 (181.5)	416.4 (431.7)	0.11	218.8 (219.2)	465.5 (454.0)	0.008
Median (Min, Max)	203.0 (9.0, 635.0)	349.0 (17.0, 2394)		156.0 (15.0, 584.0)	363.0 (9.0, 2394)		136.0 (9.0, 692.0)	378.0 (66.0, 2394)	.
Maximum ALT (U/L)									
Mean (SD)	212.1 (106.0)	343.6 (190.1)	0.02	210.6 (123.3)	323.8 (182.7)	0.10	238.6 (136.3)	335.8 (190.1)	0.10
Median (Min, Max)	208.0 (42.0, 461.0)	348.0 (69.0, 774.0)		229.0 (42.0, 461.0)	253.0 (86.0, 774.0)		229.0 (42.0, 486.0)	274.0 (69.0, 774.0)	