THE FORUM FOR COLLABORATIVE RESEARCH

LIVER FORUM HISTOLOGY SERIES

Session 1: Best Practices for Increasing Reliability

Webinar Summary
September 1, 2020

1608 Rhode Island Avenue NW
Washington, DC 20036
www.forumresearch.org

THE FORUM
For Collaborative Research℠

Berkeley Public Health

BEST PRACTICES FOR INCREASING RELIABILITY

Setting the Stage: Review of Issues and Recent Data

Presenter: Stephen Harrison, Oxford University

Slides: https://forumresearch.org/storage/documents/01_SHarrison.pdf

Background

- The level of discordance between pathologists and the differences in operationalizing liver biopsy collection in trial protocols have been raised as major issues and areas of concern for NASH clinical trials
- The goal of this meeting will be to discuss the different approaches regarding how biopsies are read, when they are read, and how training standards can be implemented.

Case Illustrations

- Three NASH patients that were screened and re-screened for clinical trials at the same site demonstrate issues with inter-rater reliability.
    - Patient #1
        - First Screening: NAS = 3 (Steatosis 1, Ballooning 0, Lobular Inflammation 2), and Fibrosis 2
            - Patient screen failed due to absence of ballooning.
            - Investigator re-screened patient for another trial using the same criteria.
        - Second read: NAS = 4 (Steatosis 1, Ballooning 1, Lobular Inflammation 2), and Fibrosis 3
            - Patient met enrollment criteria.
    - Patient #2
        - First Screening: NAS = 4 (Steatosis 2, Ballooning 0, Lobular Inflammation 2), and Fibrosis 2
            - Patient screen failed and was subsequently rescreened
        - Second Screening: NAS = 5 (Steatosis 2, Ballooning 1, Lobular Inflammation 2), and Fibrosis 2
            - Everything similar except the presence of ballooning
    - Patient #3
        - First Screening: NAS = 3 (Steatosis 2, Ballooning 0, Lobular Inflammation 1), Fibrosis 0
            - Patient screen failure and re-screened for another trial using the same criteria.
        - Second Screening: NAS = 4 (Steatosis 2, Ballooning 1, Lobular Inflammation 1), Fibrosis 2

Foundational Principles

- Surrogate Endpoints:
    - The currently accepted surrogate endpoints for regulatory approval are:
        - Resolution of NASH without worsening of fibrosis
          - or (FDA) / and (EMA) -
        - Improvement in fibrosis by at least one stage without worsening of NASH
- Fibrosis Staging
    - The pathological interpretation of fibrosis looks for architectural changes, bridging, cirrhosis, and describes where fibrosis is located.

  - There are now ways to quantify fibrosis, such as Sirius Red staining or fully quantitative assessment of collagen
  - A recent meta-analysis published by Taylor et al in Gastroenterology[1] looked at fibrosis stage and liver related outcomes
    - 13 different studies, 4,428 patients with NAFLD
    - Suggests a link between liver-related outcomes and fibrosis stage
- NASH Resolution
  - There continues to be a debate around NASH resolution, including recently at the 2020 EASL International Liver Congress
  - Liver-related mortality has been linked to the presence of NASH, with a hazard ratio of 6.28
  - The presence of NASH has also been linked to survival free of liver transplantation
- Biopsy Reliability
  - The reliability of liver biopsy in randomized clinical trials was examined in a recently published article in the Journal of Hepatology[2].
    - 339 pairs of liver biopsies that were randomized from the EMMINENCE trial (phase 2b study of MSDC-0602K) that was presented last year at AASLD.
    - The paired biopsies were independently read by three hepatopathologists that were blinded to the treatment code
    - Inter-observer kappas: steatosis = 0.69, fibrosis = 0.48, lobular inflammation = 0.3, and ballooning = 0.5
      - Whether SAF or NASH CRN criteria are used, ballooning and inflammation are defining activity and are two critical components in both scoring systems
    - Looking more specifically at the two surrogate endpoints that are currently accepted for accelerated approval:
      - Resolution of NASH with no worsening of fibrosis, unweighted Kappa = 0.396
      - Improvement of fibrosis with no worsening of NASH, unweighted Kappa = 0.366
      - Endpoints for clinical trials need to be specific, measurable, attainable, relevant, and time-bound. These results raise concerns about the measurability or repeatability of these endpoints.
        - 46% of patients that had been included in the study were deemed to not meet the criteria for study enrollment by at least one of the three other pathologists.
      - The lack of reliability of endpoints ultimately has a role in diminishing the power of a study from over 90% to as low as 40%
- Fully automated fibrosis quantification
  - Automated quantification methods are available though still experimental and being evaluated relative to histopathology[3],[4]
  - Potential to provide better accuracy between early stages of fibrosis and with possibly less sensitivity to sampling error.

[1] Taylor et al. Association Between Fibrosis Stage and Outcomes of Patients With Nonalcoholic Fatty Liver Disease: A Systematic Review and Meta-Analysis. Gastroenterology. 2020;158(6):1611-25.

[2] Davison et al. Suboptimal Reliability of Liver Biopsy Evaluation has Implications for Randomized Clinical Trials. Journal of Hepatology. 2020. In-Press.

[3] Liu et al. qFIBS: An Automated Technique for Quantitative Evaluation of Fibrosis, Inflammation, Ballooning, and Steatosis in Patients With Nonalcoholic Steatohepatitis. Hepatology. 2020;71(6):1953-66.

[4] Wang et al. Quantifying and Monitoring Fibrosis in Non-Alcoholic Fatty Liver Disease Using Dual-Photon Microscopy. Gut. 2020;69(6):1116-26.

Panel and Group Discussion

Slides: https://forumresearch.org/storage/documents/02_PanelQs.pdf

Screen Failures and Scoring Systems

Q: Industry, physicians, and patients are all impacted by the high screen failure rate with biopsy. The difficulty with identifying ballooning is particularly evident. Is it time to modify NASH CRN staging?

- The NASH CRN scoring system was developed as a method to measure the disease. In any stage-based system, there are always going to be cases that fall close to the dividing line between stages, though this is less likely to occur with more advanced disease.
- The focus is often on the biopsy resulting in screen failures; however, by nature of the trial design, the biopsy is always performed last after trial enrollment criteria have been applied. The pre-biopsy screen process also results in screen-failure of patients.
    - Developing very specific enrollment criteria will also generally increase screen failures (i.e., only patients with bridging, or, only patients with cirrhosis).
    - Regardless of the scoring system used, this problem will persist as long as there is histologic entry criteria; however, image analysis may be able to address this issue.

Q: What steps can we take to decrease screen failures?

- Data seems to support harmonization between pathologists before a trial starts. Two pathologists may be better than one if they can harmonize through a series of slides and agree on what is considered ballooning. If there is a case where a patient does not meet criteria due to ballooning or other borderline criteria, that slide could be read by a second pathologist. Where there is disagreement, the pathologists could look at it together.
- It can be confusing for patients who are close to the borderline and hear conflicting information about their disease state.
- In general, the more observations are made, the more likely to have an accurate result.
- In the PIVENS study, when the local pathologist determined that a patient fell on the threshold, that biopsy was reviewed by one or two other pathologists to ensure the patient met the criteria.
    - Methods of harmonization have been used, and increasing the number of observations can reduce the noise (but will not completely eliminate noise)
    - Limitations including keeping to the time-frame of the study, and the logistics of pathologists located in multiple locations

Q: If fibrosis is the main culprit of screening failure, why don't radiological and serological tests help minimize screening failure before a patient is admitted into a trial?

- Pre-screening strategy has significantly reduced screen-failure rate on diagnosis of NASH and fibrosis (biopsy screen-fail 30%, down from 60%)
- There are not yet reliable imaging or wet-biomarkers to detect subtle changes in fibrosis.
- The low kappa on inflammation and resolution of NASH is concerning and would require a huge signal from a drug to overcome.
    - Uncertain at this point what happens to inflammation with effective treatment, especially in the early-response phase. The time course of response by mechanism of action is not yet known.

C: The discordance rate for the NASH CRN is relatively stable across different studies; however, non-NASH CRN clinical trials seem to have a greater degree of variability.

- If the discordance rate is consistent, this can be accounted for in the sample size estimation ahead of time.
- Is there a best practice for how the biopsies should be read (pairs, masked at the end, mixed up) which could aid investigators and sponsors when setting up trial protocols?
    - The pathologists in the NASH CRN have been working together for many years which helps with the reproducibility.

Q: The NASH CRN scoring system was developed before clinical trials were being designed for NASH. Does the system meet the requirements for what is needed in a trial? Does it need to be revised according to what can be delivered in clinical trials, considering global trials and the competition for patient recruitment?

- NASH CRN is very useful, but it may be time to consider version 2.0 and re-discuss and describe borderline cases, developing a more accurate definition for the components of the scoring system that have high inter-observer variability.
    - If another scoring system is proposed, it will need to be validated and assessed for inter- and intra-observer variability, and compared to the existing system.
        - More likely to update the current system with more detailed definitions.
- Important to remember there are also studies showing high concordance and reliability of assessing features such as steatosis and fibrosis.
- Borderline cases are the source of much inter-observer variability, and at least partly related to the accuracy of the definition of the scoring.
    - Fibrosis reliability is relatively high: definitions of fibrosis stage are very accurate and there is special staining for fibrosis, and thus there are few borderline cases.
    - Ballooning and inflammation have much greater degree of inter-observer variability, which is at least partly related to the definition of each grade of the scoring system.
- The definitions of the parameters must be succinctly and clearly defined (i.e. what is a ballooned cell, what constitutes lobular inflammation)
- A set of definition criteria that could be selected and applied for particular studies will be very helpful. For example, defining ballooning as require at least 2 balloon cells in two different lobules, with a specific definition for balloon cells (i.e., twice size of normal hepatocytes, presence of Mallory bodies).
    - The trade-off of this approach would be while increasing the agreement between pathologists, the number of patients that would fit the more specified criteria and be eligible for enrollment would decrease.

Q: Would negative staining for CK-18 improve specificity for identifying a balloon cell?

- Generally this is not very useful because it identifies the typical balloon cells that do not create problems for pathologists. Possibly other antibodies that are more specific for balloon cells could help.

Q: If the patient has fibrosis, but no inflammation or no ballooning, is it right to say the patient does not have NASH?

- Similar to diabetes or hypertension, once a patient has NASH, they never really get rid of it. This brings up issues with the definition and concept of NASH resolution. The disease can be treated so that features are diminished to the point where they may not be detected.
    - The assumption of NASH resolution is that patients who have received treatment and ‘resolved’ NASH are the same as patients who never had NASH. This is unknown and a confusing point.
        - From a different perspective, the assumption is that patients with ‘resolved’ NASH had an improvement in their liver condition that makes them better off at the end of therapy than at the beginning.
        - Resolution of NASH is massive improvement in activity of the disease to the extent that the distinguishing features of activity (ballooning, inflammation) are no longer seen.
    - From a pathologist point of view, resolution of NASH is the most challenging histological parameter to assess confidently
        - The term "NASH resolution" is also confusing from a clinical and patient perspective- if a doctor tells the patient that their follow-up biopsy shows NASH resolution despite significant fibrosis, this word choice sounds very much like they have been cured.
    - Instead of using the unclear definition of ‘resolution of NASH’ as an outcome, is it worth considering if the treatment significantly alleviates the features of the disease?

- Instead of NASH Resolution, should we consider using 2-points improvement in NAS (without worsening fibrosis) as the endpoint?
    - Eliminating NASH resolution as an endpoint does not resolve the issue of inter/intra-observer variability if NAS were to be used instead, because here, with NASH being a composite score of the steatosis, ballooning and lobular inflammation scores, the inter/intra-observer variability will still affect the score each of these histological features/components
    - Defining resolution of NASH taking into account the overall histologic pattern of injury "presence or absence of definite NASH" as opposed to limiting the definition to just two of the NAS components (inflammation and ballooning) should be discussed by the pathologists. Especially taking into account kappa / concordance of inflammation and ballooning grading compared to that associated with presence or absence of definite NASH.
- Longitudinal data has shown that changes in NAS (likely also the case with SAF) over time are linked to corresponding changes in fibrosis. The degree of improvement or worsening varies according to the amount of change- e.g., a 1-point worsening is not as bad as a 3-point worsening.
    - Natural history data and data from patients involved in the NASH CRN clinical trials.
    - In the natural history of the disease, features tend to improve/ worsen simultaneously. Whether that holds for particular treatments or interventions is still unknown and possibly a treatment may improve one feature and not others.
    - Regardless of where the patient starts on the scale, a 2 or 3 point improvement or worsening of NAS will show a difference compared with a patient whose condition does not change at all.

Technical Considerations

Q: How can the technical aspects of the biopsy be improved? More training, harmonization, standardization, stricter criteria?

- The technical quality of histology seen in some trials is highly variable: shredded sections, overstained trichrome, knife-lines, chatter
    - These technical issues can influence the interpretation and add to the noise
    - Focusing on these issues in the pre-analytic phase could be an ideal place to start addressing quality in order to decrease the noise.
- It is important during the trial design phase to have a discussion of the histologic definitions for a particular study, as well as the use of a central laboratory to process biopsies, and how to ensure the technical quality of staining.
    - If pathologists can agree on how to approach borderline cases at the beginning of a study, there may be more uniformity.
- One concern about implementing very strict study criteria is that may diminish the generalizability/ clinical applicability – have to weigh this consideration
- Another option to increase intra/inter-rater reliability is to incorporate continual feedback and notes about scoring discrepancies (borderline, bad slide, etc).
- Biopsy still gives the most useful and integrated information to assess the impact of a drug. It can be complimentary to imaging and biomarkers to more fully understand about what is going on with a patient or what the impact of a treatment is.

Q: Should there be good practice guidelines in place as to how to collect an adequate liver biopsy sample for histological assessment? What is adequate length and number of portal tracts, 1 vs 2 cores, right vs left lobe, which cut of biopsy vs use of original slides used to local read, what parameters for adequate staining?

- In practice, finding that local reads are very different from central reads after additional cuts are made. Need to optimize the sample that is given to pathologists to be able to increase the yield of the biopsy.
- Having enough liver tissue to process, in addition to the quality of the section, and the quality of the stain are all very important. Particularly with fatty liver disease, an expert histotechnologist is needed to produce reliably good sections.
- Any practice guidance developed should also include industry experts in the process, to ensure what is requested can be delivered in a trial.

Pathologist Approaches
C: For clinical trials in oncology, slides have to be centrally read and multiple pathologists must concur with the diagnosis before the patient enters the trial. Two pathologists read the slides and if there is a discrepancy, a third pathologist will read it independently. If two of the pathologists agree, that will become the standard criteria. Is it possible for this model to be used in NASH trials?

- The two-pathologist model is being used for some trials and there are pros and cons to each approach. Using multiple pathologists increases data points (pro), it also increases the amount of time it takes to receive results (con).
    - Increased observations can reduce noise, but they can also potentially increase the time it takes for a patient being screened to go from signing a consent form to being randomized into the trial.
        - Consider accepting a longer screening window, and/or, consider if digital slides would function as sufficient parameters to qualify a patient for a study.
            - If digital slides could be used, they could be sent to two pathologists simultaneously, which would decrease the time needed to render an opinion and harmonize the results.
            - Digitization is a great tool to reduce variability- the technology exists and it could easily be done. One benefit is the ability to highlight histological features such as ballooning to be able to identify the cells.
            - For a large Phase 3 study with 1000s of slides, maintaining the chain of custody and reducing operational complexity are critical. The slides cannot be sent back and forth between reviewers who are located across the Atlantic. Can digital slides help, particularly with enrollment consensus? Most pathologists appear not to like to use digital slides for final central review.

C: Instead of trying to have two pathologists (or three) try to reach consensus on each slide, a different approach would be to have two separate pathologists who both have high intra-rater reliability to read and score the biopsies independently, resulting in two separate data sets. If the statistical evaluation of the datasets match (e.g., the trial has met the endpoint), this would be very robust evidence.

- If the two data sets result in disagreement (e.g., only one determines the endpoint has been met) it means the drug probably does not have a significant effect.
- The approach is not comparing each biopsy between readers, or trying to reach a consensus between readers, but rather analyzing the whole set of biopsies.
- The process for reaching a consensus for biopsy reading can result in following a ‘leader’, or, can lead to bargaining on biopsy after biopsy resulting in a data set with high variability.
- To assess the reliability of the biopsy reader, a small study could easily be done before the trial starts to determine kappa statistics. This would be a better approach than waiting until after the trial and realizing there is a problem.

Q: The biopsy is what is currently being used and is what is currently accepted by regulators – should we change the lens through which biopsy data is interpreted? How can we make better sense of the data considering the amount of discordance?

- While enhancing sample size may help obtain a significant p-value, it still does not address a diluted effect size. A significant p-value alone may not mean much if the benefit-risk is not at a desired level.
- Having two pathologists with a third as a tie-breaker sounds like a good process; however, if the intent of having a tiebreaking pathologist is that they are the more qualified expert, then a better process would be only to utilize that single pathologist.
- With NASH, there are so many parameters to assess (inflammation, ballooning, steatosis, fibrosis), and multiple permutations, that there will be occurrences where none of the pathologists align.
    - The histologic variability is not quite that severe- generally there is agreement that cirrhosis is present or there is no fibrosis. There is noise, but it is manageable and there have been a lot of studies that have shown good correlations. The methodology is fundamentally reliable- experienced observers and a consensus between observers would reduce the noise.

Q: Should baseline biopsies be re-read in a blinded fashion by the same pathologist with the week 72 biopsy, in order to reduce intra-observer variability in assessment of that patient's response?

- Yes, though will have to acknowledge that some of the baseline biopsies will not meet entry criteria, which occurred in the NASH CRN trials.
- There are other ways of looking at paired biopsies such as a blinded analysis to assess them as better, worse, or same
    - This type of analysis helps to capture intra-stage or intra-grade improvement or worsening that may not be captured by the score.
    - Such an approach has previously been done in HCV trials
    - While may be useful clinically, what is critical is whether it is acceptable from a regulatory point of view.
    - Potentially the approaches can be complimentary, and could be used to validate data. For example, looking at biopsies that were scored ‘better’ and seeing what that improvement correlate with in terms of fibrosis change, or long term outcomes.

Non-Invasive Assessment

Q: Would composite evaluation using multiple non-invasive wet biomarkers and/or histopathological endpoints improve evaluation?

- The more information collected and available to evaluation what is happening to the patients in the trial, the better what is really going on with the disease and response to an intervention can be understood.
- If a trial uses a composite endpoint, considerations such as ‘what happens if the results disagree with the primary endpoint’, or, ‘what happens if the histology looks better, but the other biomarker didn’t change’ need to be decided ahead of time.

Q: Should we try to replace biopsies with non-invasive biomarkers (which would be a surrogate of a surrogate), or should we search for biomarkers as surrogate for long-term outcome events?

- A surrogate of a surrogate is not acceptable. Currently, there is not a marker that has been proven to be related to clinical outcomes.
- While there are good non-invasive biomarkers for steatosis and fibrosis, biomarkers for disease activity are limited.
- Several non-invasive tools, specifically wet biomarkers, have been constructed with their algorithms modeled on liver biopsy. It will not be possible to beat the biopsy with biomarkers that have been constructed this way.
- If the decision to approve a drug is based on a biomarker, need to know what happens to those patients once the biomarker is reduced to an acceptable level. For example, if stiffness is reduced, once the patient is taken off therapy, how quickly is stillness reacquired?
    - The same types of questions also apply to biopsy as well- if an intervention reduces ballooning or inflammation, how quickly does this reoccur once off therapy? This type of information is important to understood in order to take care of patients.

- Clinical trials enrolling across the different mechanisms of action have been able to accumulate non-invasive data relative to histopathology. Drug developers are encouraged to continue to include non-invasive tests in their clinical trials. Even when the trial may be unsuccessful, the data generated can be very useful, particularly if the non-invasive tests can be related to clinical outcomes.
    - This data is being generated through the NIMBLE[5] and LITMUS[6].

Q: Considering the variable nature of disease activity, is the biopsy the best tool to assess activity, or should we focus more on non-invasive biomarkers?

- The concept of activity of the disease needs to be better discussed – there is clear directionality between activity of the disease and fibrosis progression or regression. Studies from NASH CRN have shown this- PIVENS[7], FLINT[8], and recent JAMA paper[9].
- Whether the parameters that can be measured in a biopsy sample are the most accurate predictor of what is happening clinically is unclear, but it is what currently can be measured.
- There are non-invasive measures available, but they are mostly measuring something else and depend on the fact that disease activity and fibrosis tend to track together.
- Over time the field has evolved and studies have been done which confirm that the histological components of activity relate to the progression or regression of fibrosis. So far this has not been done for non-invasive measures.
    - If a non-invasive biomarker is going to be used as a surrogate activity score, it must be able to measure how quickly the disease is progressing (or regressing) – studies that look at rate of progression are needed.
- Regulatory agencies fully understand this is an evolving science and remain open to hearing other options for biomarkers and surrogate endpoints, and different ways to look at histology.
    - The field must try to improve the reliability and reproducibility of biopsy as much as possible so that eventually non-invasive methods can be used for diagnosis, staging, and treatment.

Q: Could biomarkers be used along with liver biopsy in trial design and interpretation to test if the biopsy results are supported by the results of biomarkers that can indirectly confirm a beneficial effect on outcomes?

- Regulatory agencies look at the big picture, not only the liver biopsy. This includes biomarkers, and assess if results are consistent and showing improvement in the disease.
- Ideally liver biopsy would be replaced by biomarkers, but the evidence is not there yet.
- Which parameters are being used to quantify the effects of drug therapy could be more formally stated in the design of the studies and interpretation of the results.
    - Sponsors choose how to design the trial and can use key secondary endpoints that are controlled for multiplicity and other controls to ensure statistical validity.
    - Most sponsors look at multiple exploratory endpoints- regulatory agencies like to see a trend that the measures track together as evidence that the patient is getting better.
    - The same biomarker cannot be used for every drug because it depends on the drug mechanism of action and if the particular biomarker is in the pathway.

[5] Non-Invasive Biomarkers of Metabolic Liver Disease (NIMBLE). https://fnih.org/nimble

[6] Liver Investigation: Testing Marker Utility in Steatohepatitis (LITMUS). https://litmus-project.eu

[7] Brunt et al. Improvements in Histologic Features and Diagnosis Associated With Improvement in Fibrosis in Nonalcoholic Steatohepatitis: Results From the Nonalcoholic Steatohepatitis Clinical Research Network Treatment Trials. Hepatology. 2018;70(2):522-31.

[8] Neuschwander-Tetri et al. Farnesoid X Nuclear Receptor Ligand Obeticholic Acid For Non-Cirrhotic, Non-Alcoholic Steatohepatitis (FLINT): A Multicentre, Randomised, Placebo-Controlled Trial. Lancet. 2015;385(9972):956-65.

[9] Kleiner et al. Association of Histologic Disease Activity With Progression of Nonalcoholic Fatty Liver Disease. JAMA Network Open. 2019;2(10):e1912565.