NOVARTIS

Clinical Development

LJN452/LIK066

Clinical Trial Protocol CLJN452D12201C / NCT04065841

A randomized, double-blind, parallel-group, multicenter study to assess efficacy, safety, and tolerability of oral tropifexor (LJN452) & licogliflozin (LIK066) combination therapy, compared to each monotherapy, for treatment of adult patients with nonalcoholic steatohepatitis (NASH) and liver fibrosis (ELIVATE)

Statistical Analysis Plan (SAP)

Author: Statistician, [redacted]

Document type: SAP Documentation

Document status: Final 1.0

Release date: 09-Dec-2019

Number of pages: 47

Document History – Changes compared to previous final version of SAP

Date	Time point	Reason for update	Outcome for update	Section and title impacted (Current)
09-Dec-2019	Prior to FPFV	Creation of final version	N/A - First version	NA

Table of contents

List of abbreviations

ACR	Albumin-creatinine ratio
AE	Adverse event
ALP	Alkaline phosphatase
ALT	Alanine transaminase
ANCOVA	Analysis of covariance
[redacted]	[redacted]
APTT	Activated partial thromboplastin time
AST	Aspartate transaminase
ATC	Anatomical Therapeutic Classification
BMI	Body mass index
[redacted]	[redacted]
[redacted]	[redacted]
CI	Confidence interval
[redacted]	[redacted]
[redacted]	[redacted]
CSR	Clinical study report
CTCAE	Common Terminology Criteria for Adverse Events
CV	Coefficient of variation
DMC	Data monitoring committee
[redacted]	[redacted]
eCRF	Electronic case report form
eGFR	Estimated glomerular filtration rate
[redacted]	[redacted]
EOS	End of study
EOT	End of treatment
FAS	Full analysis Set
FGF	Fibroblast growth factor
[redacted]	[redacted]
GGT	Gamma glutamate transaminase
HA	Hyaluronic acid
[redacted]	[redacted]
HDL-C	High density lipoprotein cholesterol
[redacted]	[redacted]
IL	Interleukin
INR	International normalized ratio
ITT	Intent-to-treat
LDL-C	Low density lipoprotein cholesterol
LLOQ	Lower limit of quantification
MAR	Missing at random
MedDRA	Medical dictionary for regulatory activities
MMP	Matrix metalloproteinase

MMRM	Mixed model repeated measures
MRI	Magnetic resonance imaging
NAFLD	Non-alcoholic fatty liver disease
NAS	NAFLD activity score
NASH	Non-alcoholic steatohepatitis
OCA	Obeticholic acid
PCR	Protein-creatinine ratio
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
[redacted]	[redacted]
PT	Preferred term / Prothrombin time
RAN	Randomized set
RNA	Ribonucleic acid
SAE	Serious adverse event
SAF	Safety set
SAP	Statistical analysis plan
SCR	Screened analysis set
SD	Standard deviation
SMQ	Standardized MedDRA Query
SOC	System organ class
TBL	Total bilirubin
TD/EOT	Treatment discontinuation / End of treatment
TFLs	Tables, figures, listings
TIMP	Tissue inhibitors of metalloproteinases
[redacted]	[redacted]
TT	Thrombin Time
ULN	Upper limit of normal range
ULOQ	Upper limit of quantification
VAS	Visual analogue scale
VLDL-C	Very low density lipoprotein cholesterol
WHO	World Health Organization

1 Introduction

The purpose of this Statistical Analysis Plan (SAP) is to provide details on the implementation of analyses outlined in section 12 of the study protocol and to be reported in the core Clinical Study Report (CSR).

Analyses based on this SAP will be executed after all patients complete the study.

Data analyses for Data Monitoring Committee (DMC) meetings will be specified in a separate DMC SAP.

[redacted]

1.1 Study design

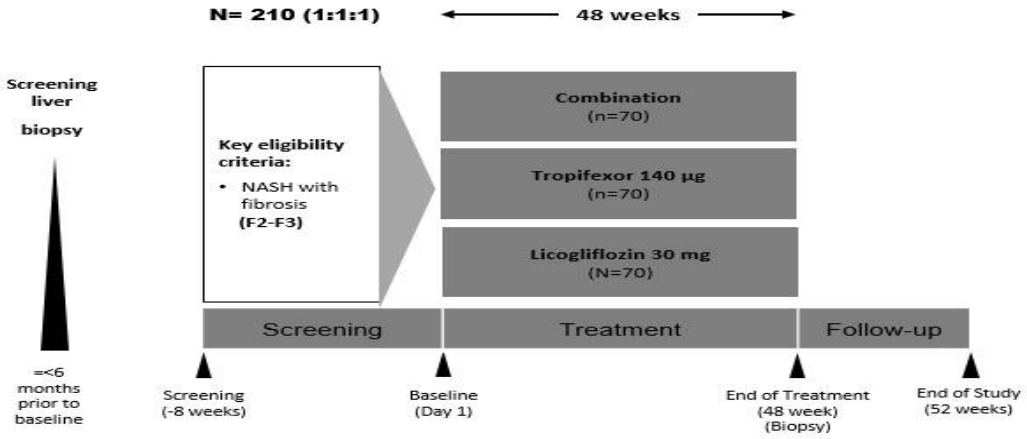
This is a randomized, double-blind, parallel-group, multiple-arm study to assess the efficacy, safety and tolerability of tropifexor and licogliflozin combination therapy compared to each monotherapy in patients with non-alcoholic steatohepatitis (NASH) and advanced fibrosis.

The study consists of 1) a screening period, 2) a treatment period starting from randomization on Day 1 and running to Week 48, and 3) a follow-up period of 4 weeks after the last dose of study treatment. The screening period starts from the time of the signing informed consent and continues for up to 8 weeks when all inclusion/exclusion criteria have been evaluated and all baseline assessments have been performed. The study duration from first dose of study medication is 52 weeks. The total duration of participation may be up to 60 weeks.

Approximately 210 patients with NASH and F2-F3 fibrosis will be randomized in a 1:1:1 ratio to one of the following treatments:

- Arm A: tropifexor + licogliflozin combination therapy,
- Arm B: tropifexor monotherapy,
- Arm C: licogliflozin monotherapy

Figure 1-1 Study Design



The primary endpoint data for efficacy is collected at Week 48.

In case of any protocol amendment, this SAP may be amended accordingly.

1.2 Study objectives and endpoints

Table 1-1 Objectives and related endpoints

Objective(s)	Endpoint(s)
Primary objective(s)	**Endpoint(s) for primary objective(s)**
• To demonstrate the efficacy of tropifexor + licogliflozin as assessed by histologic improvement after 48 weeks of combination treatment compared to each monotherapy treatment in patients with NASH and stage 2 or 3 fibrosis.	Proportion of patients with resolution of NASH and no worsening of fibrosis OR improvement in fibrosis by at least one stage without worsening of NASH (no worsening of hepatocellular ballooning or lobular inflammation) at Week 48 compared with baseline. The criteria for resolution of NASH (absence of ballooning with no or minimal inflammation by histology) and other changes in histologic endpoints are further detailed in Section 5.3.11.
Secondary objective(s)	**Endpoint(s) for secondary objective(s)**
• Improvement in fibrosis by at least one stage with no worsening of NASH after 48 weeks of treatment	• **Key secondary endpoint:** Proportion of patients who have at least one stage improvement in fibrosis without worsening of NASH (no worsening of hepatocellular ballooning or lobular inflammation) at Week 48 compared with baseline

Objective(s)	Endpoint(s)
• Resolution of NASH with no worsening of fibrosis after 48 weeks of treatment	• **Key secondary endpoint:** Proportion of patients with resolution of NASH and no worsening of fibrosis at Week 48 compared with baseline
• Improvement in fibrosis by at least one stage	• Proportion of patients who have at least one stage improvement in fibrosis at Week 48 compared with baseline
• Improvement in fibrosis by at least two stages with no worsening of NASH after 48 weeks of treatment	• Proportion of patients who have at least two stage improvement in fibrosis without worsening of NASH (no worsening of hepatocellular ballooning or lobular inflammation) at Week 48 compared with baseline
• Reduction in body weight from baseline after 48 weeks of treatment	• Proportion of patients with 5% or more reduction in body weight at Week 48 compared to baseline
• Change in liver fat content after 48 weeks of treatment	• Change in liver fat content based on MRI - PDFF over time up to Week 48 compared with baseline
• To determine the relationship of investigational treatment and markers of hepatic inflammation in NASH (ALT and AST)	• Change in ALT and AST over time up to Week 48 compared with baseline
• To determine the relationship of investigational treatment and GGT, a marker of cholestasis and oxidative stress	• Change in GGT over time up to Week 48 compared with baseline
• To evaluate the safety and tolerability of tropifexor (LJN452) in combination with licogliflozin (LIK066), compared to monotherapy of each compound from baseline	• Occurrence of adverse events, serious adverse events, adverse events resulting in discontinuation of study treatment, adverse events of special interest and changes in vital signs and laboratory tests

2 Statistical methods

2.1 Data analysis general information

This SAP guides the statistical analysis for the core CSR after all patients complete (including early withdrawal) the study. There is no planned interim analysis for statistical inference. DMC analyses are described in a separate SAP.

Novartis statisticians and programmers will perform this statistical analysis, primarily using SAS 9.4 or newer. Additional analysis software (e.g. R), in their newest versions available in the validated data analysis environment, may also be used.

Descriptive summaries will be presented by study treatment without a total column, unless requested in a specific section. Categorical variables will be summarized with count and percentages. If not otherwise specified, continuous variables will be summarized with n, arithmetic mean, standard deviation, minimum, maximum, median, first quartile and third quartile.

Confidence intervals (CI) in this SAP refers to 2-sided confidence intervals unless clarified differently. P-values will be reported only if a formal hypothesis test is performed.

Analyses performed at pre-specified analysis visits, including the primary efficacy endpoint at week 48, will be based on the analysis visit instead of original CRF visits. The derivation of analysis visits is detailed in Section 2.1.1. As an example, if the scheduled visit 5 (W4) of a patient is delayed and occurs on study day 48 instead of on Day 28, the data from that visit will be mapped to analysis visit 6 (W8).

Analysis by visit (or "over time"**)** will present baseline and only post-baseline visits at which an assessment for the endpoint is scheduled according to protocol (Protocol Table 8-1). Summaries at post-baseline visit will include only patients with non-missing baseline and at least one non-missing post baseline results for the endpoint.

Data listings will be presented by study treatment and include, in addition to the columns explicitly requested in specific section, following columns: patient ID, age/gender/race, first/last dose date of study treatment and date (study day) of assessment/event/sampling. All listings will present data as collected without any imputation needed for other data analysis purpose.

2.1.1 General definitions

Study treatment in this SAP refers to the investigational regimens, i.e. the combination of tropifexor (LJN452) 140 µg and licogliflozin (LIK066) 30 mg, or each of them as monotherapies combined with matching placebo for the other. For simplicity hereafter, "LJN452" and "LIK066" in this SAP refer to the two specific dose strength of LJN452 140 µg and LIK066 30 mg unless otherwise specified.

Date of first study treatment ("first dose date" or "FDDT") is the date of first administration of study treatment during the study as recorded in eCRF (Dosage Administration Record summary page).

Date of last study treatment ("last dose date" or "LDDT") is the date of last administration of study treatment during the study as recorded in eCRF (Dosage Administration Record summary page).

Study day is number of days counted from the first dose date (Day 1).

For dates on or after FDDT, study day = date of interest – FDDT + 1.

For dates prior to FDDT, study day = date of interest – FDDT.

If first dose date is missing, the randomization date will instead be used for the calculation.

Baseline for analysis, in general, is the last assessment with valid data before the date and time of first administration of study treatment. If the date of assessment is the same as first dose date but time is not available for comparison, the assessment is assumed to be prior to the administration of study treatment. For ALT, AST, ALP, GGT and total bilirubin, however, the baseline is the average of values from two screening visits and the baseline visit collected before first administration of study treatment. If there is unscheduled visit with lab assessment after a planned screening visit, the value from this unscheduled visit will also be used in the calculation of baseline.

End of study (EoS) or last contact date of a patient is the date of his/her last office visit or other route of follow-up in the study.

Treatment period is the period from the Baseline visit (included) to the TD/EOT (W48) visit (included) as defined in protocol Table 8-1.

Follow-up period is the period from TD/EOT (W48) visit (not included) to EoS (included).

Analysis visit is the visit label derived based on study day of assessment and the analysis visit windows defined in Table 2-1. For laboratory and bioanalytical parameters specified for both Group 1 and Group 2 in Table 2-1, two sets of analysis visits will be derived and will be referred to as "**AV1**" and "**AV2**" respectively. The AV1 is derived according to Group 1 for by visit analysis on full analysis set (FAS). For AV2, assessments >14 days after the last dose date will be slotted into the analysis visit window for follow up visit for by-visit analysis of laboratory data on safety analysis set (SAF).

Analysis visit windows are non-overlapping periods each associated with a visit planned in protocol and all together cover the entire study period of a patient. If multiple measurements of same parameter for a patient fall in the same window, the value of this analysis visit is selected according to Table 2-2. If no assessment falls into a window, the value of that analysis visit will be set to missing unless imputation rule(s) are defined for the parameter. Repeat and unscheduled visits will be mapped to analysis visits in the same way as scheduled visits.

Table 2-1 Analysis visit windows

	Analysis Visit Windows

Analysis Visit	Target Study Day	Group 1	[redacted]	Group 4	Group 5
Baseline	1	≤ 1*		≤ 1*	≤ 1*
Week 2	15	2-22			
Week 4	29	23-43			
Week 8	57	44-71			
Week 12	85	72-99		2-127	2-169
Week 16	113	100-127			
Week 20	141	128-155			
Week 24	169	156-197		128-253	
Week 32	225	198 – 253			
Week 40	281	254 – 309			
Week 48	337	310 – 351		254 - 351	170 - EoS
Wk 52/FU	365#	352 – EoS		352 - EoS	NA

* The first dose date of the study is defined as Day 1. If first dose date is missing, the date of randomization will be used as day 1.

The target study day for the FU visit of Group 2 will be LDDT+28

Group 1: vital signs and anthropometric measurements, liver function tests (ALT, AST, TBL, Albumin, ALP and GGT), lipids panel, coagulation panel, [redacted] BUN, creatinine, eGFR [redacted]

Group 4: Waist/Hip circumference, [redacted]

Group 5: MRI, liver biopsy

Table 2-2 Choose value among multiple assessments in the same analysis visit window

Timing of measurement	Type of data	Value selection rules
Baseline	All data excluding ALT, AST, ALP, GGT, TBL, [redacted].	Choose the latest non-missing measurement before the first administration of study treatment (FDDT). If a patient did not receive any dose of study treatment, the latest measurement on or before the randomization date is used. If the date of a measurement is the same as FDDT but time is not available for comparison, the measurement is assumed to be prior to the first administration of study treatment.
	ALT, AST, ALP, GGT, TBL, [redacted]	The average of valid values from all visits before first administration of study treatment will be calculated as baseline.
Post-baseline	Continuous parameters	Choose the measurement closest to the target study day. If more than one measurements on the same date, choose the last (based on time, or visit label if time not available) measurement on that day.

Timing of measurement	Type of data	Value selection rules
		If two measurements are equally away before and after the target study day, use the mean.
	Categorical efficacy parameters	Choose the measurement closest to the target study day. If more than one measurements on the same date, choose the last (based on time, or visit label if time not available) measurement. If two measurements are equally away before and after the target study day, use the first measurement.
	Notable laboratory abnormalities; shift from baseline;	The most extreme measurement in the window will be used. Note that this means a patient can have a notably high and notably low measurement within a window.

Geographical regions are defined according to countries of study sites at which patients are randomized:

Table 2-3 Geographical regions

Region	Countries
Europe	Belgium, Bulgaria, Denmark, Estonia, France, Germany, Italy, Russia and CIS, Spain, United Kingdom
North America	Canada, USA,
Latin America	Argentina, Brazil, Canada, Chile, Colombia, Mexico
Asia	China, India, Japan, Singapore, South Korea, Taiwan,
Other	Australia, Egypt, Saudi Arabia, South Africa, Turkey

2.2 Analysis sets

The following analysis sets are defined for analysis purpose:

Screened analysis set (SCR) comprises of all patients who signed the informed consent. This analysis set includes also screen failures.

Randomized set (RAN) comprises of all patients who received a randomization number, regardless of whether patient received study medication. Mis-randomized (not qualified for randomization but randomized inadvertently) patients will be included in this analysis set.

Full Analysis Set (FAS) comprises of all patients to whom study treatment has been assigned by randomization. Patients who were mis-randomized (not qualified for randomization but randomized inadvertently) and did not receive investigational medication are excluded from the FAS. Efficacy analysis will be conducted on FAS. Following the intent-to-treat (ITT) principle, patients are analyzed according to the study treatment assigned during the randomization regardless of actual treatment received and according to the true stratum the patient belongs to in case a patient is assigned to wrong stratum during randomization.

Safety Analysis Set (SAF) includes all patients who received at least one dose of study treatment. Patients with the following unusual situations, if it occurs, will be excluded from SAF.

- Had no safety assessment after first dose of study treatment (e.g. lost to follow-up). Of note, the statement that a subject had no adverse events constitutes a safety assessment.
- Took only matching placebo tablet or capsules during the study. Of note, such situation will be identified only after study unblinding.

SAF will be used for analysis of Safety [redacted] in this SAP. [redacted]

In the SAF, patients will be analyzed according to the study treatment received, where treatment received is defined as the randomized/assigned study treatment if the patient took at least one dose of that treatment or the first treatment received if the randomized/assigned study treatment was never received.

2.2.1 Subgroup of interest

Exploratory analysis on the primary endpoint (Section 2.13.1) will be performed on following subset of FAS.

- FAS with baseline fibrosis stage of F2
- FAS with baseline fibrosis stage of F3
- FAS with T2DM at baseline
- FAS without T2DM at baseline
- FAS asian (Based on race collected on the Demographics CRF page)
- FAS non-asian (Based on race collected on the Demographic CRF page)

2.3 Patient disposition, demographics and other baseline characteristics

Demographics and baseline characteristics listed below will be summarized by study treatment on FAS.

- Age (years): per demographics CRF
- Age Group: two categories derived for age <65 years or age ≥65 years
- Gender: per demographics CRF
- Region: derived based on country of study site (Section 2.1)
- Ethnicity: per demographics CRF
- Race: per demographics CRF
- Height (cm) at screening visit
- Body Weight (kg)
- Waist circumference (cm)
- Hip circumference (cm)
- BMI

- Liver function (ALT, AST, ALP, GGT, TBL): see Section 2.1.1 for special baseline definition
- [redacted]
- Steatosis stage (0, 1, 2 or 3) based on liver biopsy: categorical data summary
- Lobular inflammation stage (0, 1, 2 or 3) based on liver biopsy: categorical data summary
- Hepatocyte ballooning stage (0, 1 or 2) based on liver biopsy: categorical data summary
- NAS score (Section 5.3.8) based on liver biopsy: continuous data summary
- Fibrosis stage based on liver biopsy: categorical data summary for score 2 and 3
- Hepatic fat fraction based on MRI-PDFF
- Fasting lipids (HDL-C, LDL-C, VLDL-C, triglycerides, free glycerol, free fatty acid, total cholesterol, ApoA1)
- hsCRP
- [redacted]
- T2DM status (Section 5.3.7)
- [redacted]

Relevant medical histories up to baseline, including NASH specific co-morbidities, will be summarized by system organ class, preferred term and by study treatment on FAS.

2.3.1 Patient disposition

The number of subjects in each analysis set will be presented overall for SCR and by study treatment (as randomized) for other analysis sets. Patients with the treatment received (Section 2.2 for SAF) different from the treatment group as randomized will be listed in the footnote.

The number (%) of subjects in RAN who completed the study, who withdraw from study early and the reason for early withdrawal will be presented by study treatment.

The number (%) of patients with protocol deviations and criteria leading to their exclusion from FAS and SAF will be presented for RAN.

2.4 Treatments (study treatment, rescue medication, concomitant therapies, compliance)

The analyses of treatments will use SAF.

2.4.1 Study treatment / compliance

The duration of exposure (days) to study treatment will be calculated as the LDDT – FDDT +1 and summarized by study treatment. In addition, the number (%) of patients with LDDT in the following time intervals after FDDT will be summarized by study treatment:

- < 4 weeks
- ≥4 and < 8 weeks
- ≥ 8 and < 12 weeks
- ≥ 12 and < 16 weeks
- ≥ 16 and < 20 weeks
- ≥ 20 and < 24 weeks
- ≥ 24 and < 32 weeks
- ≥ 32 and < 40 weeks
- ≥ 40 and < 48 weeks
- ≥ 48 weeks

The proportion of patients with dose reduction (including interruption) due to adverse event, as recorded in the change in dosing log CRF page, will be presented by study treatment.

Dose intensity is derived as the ratio between the cumulative dose of each active ingredient (tropifexor and licogliflozin) received during the study and duration of exposure (days). Patients in combination arm will have dose intensities derived separately for tropifexor and licogliflozin.

The relative dose intensity (compliance) is derived as the ratio between the actual dose intensity and the planned dose intensity which equals to the actual duration of exposure * the daily dose of tropifexor or licogliflozin as planned according to protocol.

Number (%) of patients incompliant with planned regimen, defined as relative dose intensity >1.25 or <0.75, will be summarized for tropifexor in the LJN452 monotherapy arm, for licogliflozin in the LIK066 monotherapy arm, and for both tropifexor and licogliflozin separately in the combination arm.

2.4.2 Prior and concomitant therapies

Medications or other significant non-drug therapies during the study will be coded using the WHO dictionary.

Prior therapies are defined as any medications or significant non-drug therapies taken prior to the first administration of study treatment, regardless of whether stopped or continued after randomization.

Concomitant medications or non-drug therapies are defined as those received during the study treatment and follow-up periods (i.e., until end of study visit) regardless of whether started before randomization. Therefore, a medication can be BOTH prior and concomitant.

Number (%) of patients using prior and concomitant medications will be summarized separately by the 3rd and 5th level ATC terms and by study treatment on FAS.

2.5 Analysis of the primary objective

The primary objective of the study is to demonstrate the efficacy of tropifexor + licogliflozin as assessed by histologic improvement after 48 weeks of combination treatment compared to each monotherapy treatment in patients with NASH and stage 2 or 3 fibrosis.

The following estimands framework is adopted to evaluate the objective.

1. Population: Full Analysis Set (FAS), for the adult population with NASH and stage 2 or 3 liver fibrosis as further defined with study inclusion and exclusion criteria.
2. Variables of interest: histological response defined as resolution of NASH (absence of ballooning with no or minimal inflammation by histology) and no worsening of fibrosis OR improvement in fibrosis by at least one stage without worsening of NASH (worsening of hepatocellular ballooning or lobular inflammation) at week 48 compared with baseline. The criteria for changes in histologic endpoints are further detailed in Section 5.3.11.
3. Summary measure: odds ratio for the proportion of patients achieving histological response at week 48 between the combination and each monotherapy treatment.
4. Intercurrent events include
   - Early discontinuation of study treatment or withdrawal from study before week 48 visit.
   - Non-histologic clinical event suggesting treatment failures (e.g progression to cirrhosis, liver transplant and death due to disease).
   - Study treatment incompliance defined as relative dose intensity >1.25 or <0.75.
   - More than four weeks of accumulative exposure to any prohibited medications (Protocol Section 6.2.2), alcohol usage more than what is permitted in the protocol, initiation of other T2DM or lipid lowering medication (e.g. GLP-1 modulator), initiation of other treatment being evaluated or approved for NASH, or new concomitant interventions for weight control (e.g bariatric surgery) before the assessment for primary endpoint. The identification of relevant medications will be based on ATC3 and/or ATC5 codes, which will be included in analysis data set specifications. The list of concomitant interventions with significant effect for general weight loss will be determined based on blinded review of the study data.

All practical measures will be exercised in the study to collect as much data as possible at week 48 despite intercurrent events (e.g. early discontinuation of study treatment). In the case of early withdrawal from the study on or after week 24, patient's histological response will be assessed if possible at the early termination visit.

A combination of the while-on-treatment and hypothetical strategies as defined in ICH E9 (R1) on handling intercurrent events will be adopted for the primary analysis.

If a patient discontinued study treatment early or withdrew early from the study without other inter-current event, but has histological response status collected after at least 24 weeks of study treatment, the primary endpoint will be imputed with their last histological response status

assessed after at least 24 weeks of treatment. Patients with missing primary endpoint or primary endpoint after inter-current event(s) will have their primary endpoint data imputed with a multiple imputation (MI) method as outlined in Section 2.5.3 to estimate the treatment effect that would have been observed if all patients had continued on treatment for 48 weeks without intercurrent event.

2.5.1 Primary endpoint

The primary endpoint is the proportion of patients experiencing histological response, which is defined as resolution of NASH (absence of ballooning with no or minimal inflammation by histology) and no worsening of fibrosis OR improvement in fibrosis by at least one stage without worsening of NASH (worsening of hepatocellular ballooning or lobular inflammation) at week 48 compared with baseline. . The criteria for resolution of NASH (absence of ballooning with no or minimal inflammation by histology) and other changes in histologic endpoints are further detailed in Section 5.3.11.

2.5.2 Statistical hypothesis, model, and method of analysis

To demonstrate improved efficacy of tropifexor+licogliflozin in combination, the response rate in the combination arm will be compared with each monotherapy arm. A one-sided test ($\propto$=0.05) of statistical superiority is planned for each comparison. A stepwise gatekeeping hierarchical test procedure, in which the second test is only performed if the first null hypothesis is rejected, will be adopted to maximize the power for the first test while still maintaining a strong control of family wise type 1 error rate (FWER-1) at 0.05.

The hierarchical testing will be performed in the following sequence.

1. Combination vs. licogliflozin (LIK066)
2. Combination vs. tropifexor (LJN452)

The order of the hierarchical tests is determined under a preliminary assumption that tropifexor, by targeting the Farnesoid X receptor that directly modulates liver inflammation and fibrosis associated with NASH, could potentially be more effective than licogliflozin hence results in smaller add-on effect of the combination when used as the reference in the comparison. If external data newly available during the study warrants different assumptions, the order of tests could be switched before the study unblinding without inflation of the FWER-1.

For these two tests on the same endpoint and sharing a common comparator (combination arm), the hierarchical testing approach also allows an intrinsic adjustment of the correlation between them for the power of the second test.

The proportions of histological responders in the three treatment arms are assumed following binomial distributions (**n**, **P**) and the sample distribution of the difference p_1-p_{2i} will be asymptotically normal with mean P1-P2i and variance P1*(1-P1)/n1 + P2i*(1-P2i)/n2i., where

$i \in \{1, 2\}$ ~ the level of the hierarchical test

P = (P_1, P_{21}, P_{22}) ~ True probabilities of histological responses after 48 weeks of treatment with the combination, licogliflozin and tropifexor respectively

n = (n_1, n_{21}, n_{22}) ~ Number of patients in the combination, licogliflozin and tropifexor groups respectively in the study

p = (p_1, p_{21}, p_{22}) ~ Proportion of histological responders observed in the study after 48 weeks of treatment with the combination, licogliflozin and tropifexor respectively

With the sample data collected in the study, the naive differences P_1-P_{2i} can be estimated using p_1-p_{2i} with unpooled standard error SE = sqrt(p_1*(1-p_1)/n_1 + p_{2i}*(1-p_{2i})/n_{2i}).

The hierarchical testing procedure based on odds ratio can be formulated as one-sided statistical tests ($\propto$=0.05) for

$$\text{H0i} : \frac{P1(1-P2_i)}{P2(1-P1_i)} <= 1 \text{ vs. H1i} : \frac{P1(1-P2_i)}{P2(1-P1_i)} > 1$$

executed sequentially for i=(1, 2), where H02 is tested only if H01 is rejected.

A logistic regression model with the logit of probability for histological response as the dependent variable, treatment groups as the main effect, baseline fibrosis stage (2 vs. 3) and T2DM (No/Yes) status as co-factors, and baseline BMI as a covariate will be utilized to test the hypothesis at each step of the hierarchical testing procedure.

The choice of co-factors/covariates, the considerations for their collinearity, potential interactions with treatment effect and model specifications are further discussed in Section 5.4.1.

The logistic regression analyses can be implemented in SAS LOGISTIC procedure with the *scale=williams* and *firth* options in the model statement, executed together with the multiple imputation (MI) methods outlined in Section 2.5.3.

Odds ratios and 95% CI of the probabilities of histologic response between study treatments can be back calculated as the exponentiation of the corresponding coefficients estimated from the logistic regression.

One logistic regression model can be fitted on all FAS data from the three treatment arms to estimate the pairwise odds ratios and associated p-values.

Given placebo effects (Table 3-1) reported in the published literature and with Firth correction, a model convergence issue is not expected. In the case a convergence issue occurs and cannot be resolved with other appropriate technical approaches, a reduced model may be used for the primary analysis. The reduced model will be determined by removing baseline BMI, fibrosis stage and T2DM status sequentially from the full model until the convergence issue is resolved. If a reduced model has to be used, impacts on the interpretation of primary analysis results will be discussed in the core CSR in combination with results from sensitivity analysis.

Existence of extremely influential baseline BMI values will be examined with Cook's distance plot based on complete cases before MI.

The numbers and percentages of histological responders will be summarized by study treatment. The difference between the combination therapy and each monotherapy will be presented along with 95% CI calculated based on the normal approximation for distribution of the difference

between two binomial random variables (Wald asymptotic CI). Odds ratios (95% CI) and p-values from the logistic regression model will be reported.

2.5.3 Handling of missing values/censoring/discontinuations

Especially due to the clinical burden of liver biopsy, missing data is possible and could be related to observable data or even the unobserved primary endpoint.

While not possible to confirm missing completely at random (MCAR), the histological response status from the early termination visit after at least 24 weeks of study treatment will be used for the primary analysis in case of missing primary endpoint at week 48. Patients otherwise with primary endpoint missing or any primary endpoint after intercurrent events will have the values imputed using multiple imputation (MI) methods (Rubin 1976, 1987) under a missing at random (MAR) assumption.

An arbitrary missing pattern will be assumed and a fully conditional specification (FCS) method with logistic regression will be applied. Baseline fibrosis stage (2 vs. 3), baseline T2DM (No/Yes), baseline BMI (as categorical predictor) and decrease of ALT at week 24 (<= -17 U/L vs. >-17 U/L) will be included as predictors for the imputer's model. The logistic model outlined for primary analysis will be used to analyze the multiply imputed datasets, and pooled results will be obtained based on Rubin's combination rules.

More details for the MI method are discussed in Section 5.4.2.

The MI methods may be implemented using the SAS MI procedure with FCS LOGISTIC statement and the MIANALYZE procedure. The order of variable imputation will be specified with ORDER=VAR option in FCS statement. The VAR statement will list variables for baseline fibrosis stage (2 vs. 3), baseline T2DM status (No/Yes), baseline BMI (as continuous predictor) and decrease of ALT at week 24 (<= -17 U/L vs. >-17 U/L) in sequence. The number of imputations will be set to 50.

The number (%) of patients with missing primary endpoint at week 48 will be summarized by study treatment.

2.5.4 Supportive analyses

Supportive analyses will be performed to cross check with the results from primary analysis.

1. The primary analysis will be repeated with an intent-to-treat (ITT) approach, recognizing that inter-current events could happen in real clinical practice potentially as a result of the treatment policy. Patients' histological response status collected at week 48 or early termination visit will be used in this sensitivity analysis regardless of inter-current event(s). Patients otherwise with missing primary endpoint will be counted as non-responders. This approach is considered more conservative for estimating the benefit of combination therapy.
2. Tests in the two hierarchical steps of primary analysis will be repeated separately using Cochran-Mantel-Haenszel (CMH) method stratified by baseline fibrosis stage (2 vs. 3), T2DM (No/Yes) and BMI (>= 35 vs. < 35). Primary endpoint observed at week 48 or last visit after at least 24 weeks of exposure to study treatment will be used regardless of intercurrent event(s). Patients otherwise will be counted as non-

responders. More details on the CMH methods are discussed in Section 5.4.4. The CMH analysis may be implemented using SAS FREQ procedure with the CMH option in TABLES statement.

2.6 Analysis of the key secondary objective

Descriptive summaries of key secondary endpoints (Table 2-4) will be presented by study treatment on FAS. Hypothesis tests for treatment differences in key secondary endpoints will be performed if and only if both null hypothesis in the primary analysis are rejected.

2.6.1 Key secondary endpoints

Table 2-4 Key secondary endpoints and methods of analysis

Endpoints	Analysis method
Key secondary endpoint: Proportion of patients achieving at least one point improvement of fibrosis stage without worsening of NASH (no worsening of hepatocellular ballooning or lobular inflammation) at Week 48 compared with baseline	Same logistic regression model as used in the primary analysis for each step of hierarchical test
Key secondary endpoint: Proportion of patients achieving resolution of NASH without worsening of fibrosis stage at Week 48 compared with baseline	Same logistic regression model as used in the primary analysis for each step of hierarchical test

2.6.2 Statistical hypothesis, model, and method of analysis

Extending the hierarchical testing procedure for primary analysis and defining a general one-sided ($\propto$=0.05) statistical hypothesis test Hi for

$$\mathrm{H}i_0: \frac{Pi_1(1-Pi_2)}{Pi_2(1-Pi_1)} <= 1 \text{ vs. } \mathrm{H}i_1: \frac{Pi_1(1-Pi_2)}{Pi_2(1-Pi_1)} > 1$$

at each step of hypothesis test, where i=1, ..., 6 with Pi_1 and Pi_2 correspond to the true probabilities of response for combination vs. monotherapy in the ith test, the extended hierarchical testing procedure can be formalized following general principles in the graphical approach (Bretz et al 2009) to maintain a strong control of an overall type I FWER at 0.05.

where

H1: Combo vs. LIK066 comparison in primary endpoint (Section 2.5.2)

H2: Combo vs. LJN452 comparison in primary endpoint (Section 2.5.2)

H3: Combo vs. LIK066 comparison in proportion of patients achieving at least one point improvement of fibrosis stage without worsening of NASH (no worsening of hepatocellular ballooning or lobular inflammation) at Week 48 compared with baseline

H4: Combo vs. LJN452 comparison in proportion of patients achieving at least one point improvement of fibrosis stage without worsening of NASH (no worsening of hepatocellular ballooning or lobular inflammation) at Week 48 compared with baseline

H5: Combo vs. LIK066 comparison in proportion of patients achieving resolution of NASH without worsening of fibrosis stage at Week 48 compared with baseline

H6: Combo vs. LJN452 comparison in proportion of patients achieving resolution of NASH without worsening of fibrosis stage at Week 48 compared with baseline

A specific hypothesis test H*i* will be performed if and only if the null hypotheses in all previous testing steps are rejected.

The number (%) of responders will be summarized by study treatment. The difference between the combination therapy and each monotherapy will be presented along with Wald asymptotic 95% CI. The same logistic regression model and MI method for primary analysis will be adopted for hypothesis tests on key secondary endpoints and odds ratios (95% CI) will be presented. If no hypothesis test is performed for a step based on the pre-specified hierarchical testing procedure, the p-values will not be reported.

2.6.3 Handling of missing values/censoring/discontinuations

The same MI methods outlined in Section 2.5.3 for primary analysis will be used for missing data imputation during key secondary analysis.

2.7 Analysis of secondary efficacy objective(s)

Secondary efficacy endpoints in Table 2-5 will be summarized by visit and by study treatment on FAS.

2.7.1 Secondary efficacy endpoints

Table 2-5 Secondary efficacy variables and methods of analysis

Endpoint	Analysis
Proportion of patients who have at least one stage improvement in fibrosis at week 48 compared with baseline	Same analysis methods as for the primary endpoint following the ITT principle (Section 2.5.4) without hypothesis test.
Proportion of patients who have at least two stage improvement in fibrosis and no worsening of NASH (no worsening of hepatocellular ballooning or lobular inflammation) at week 48 compared with baseline	Same analysis methods as for the primary endpoint following the ITT principle (Section 2.5.4) without hypothesis test.
Change in liver fat content based on MRI - PDFF (in 60% of patients) over time up to Week 48 compared with baseline	Repeated measurement analysis of covariance (ANCOVA) model (Section 5.4.3) will be used to estimate treatment group differences in mean change from baseline and 95% CI. This analysis is on FAS patients with at least one MRI-PDFF assessment.
Proportion of patients with 5% or more reduction in body weight at Week 48 compared to baseline	CMH test stratified with baseline T2DM status will be performed to estimate odds ratio (95% CI) between combination arm and each monotherapy

2.7.2 Statistical hypothesis, model, and method of analysis

There is no hypothesis test planned for these additional secondary endpoints.

Binary histologic endpoints will be analyzed using the same logistic regression model for primary endpoint following an ITT approach (Section 2.5.4). The number (%) of responders will be summarized by study treatment. The difference between the combination therapy and each monotherapy will be presented along with Wald asymptotic 95% CI. The Odds Ratios (with 95% CI) between combination and each monotherapy estimated from the logistic regression model will also be provided.

Repeated measurements ANCOVA models for the change in liver fat fraction will include time (visit) and treatment group as categorical explanatory variables, with interaction between treatment and visit also included in the model. More details of the model are discussed in Section 5.4.3. The analysis may be implemented with SAS MIXED procedure, assuming an unstructured variance-covariance matrix for repeated measurements and using Kenward-Rogers type degrees of freedom. Estimates of the differences (95% CI) between treatment groups may be obtained using the LSMEANS and ESTIMATE statements in the SAS procedure at individual time points. Descriptive statistics, adjusted mean and adjusted mean differences (95% CI) between combination and monotherapies will be presented by visit.

For proportion of patients with at least 5% reduction of body weight at week 48, the number (%) of patients will be summarized by study treatment. Difference in percentages will be presented with Wald asymptotic 95% CI. A CMH analysis stratified by baseline T2DM status will be performed to estimate the common odds ratio (95% CI) between combination and each monotherapy. CMH analysis may be implemented with the SAS FREQ procedure following similar implementation in the sensitivity analysis on primary endpoint (Section 5.4.4) but with only baseline T2DM status as stratification factor.

2.7.3 Handling of missing values/censoring/discontinuations

The repeated measures ANCOVA (MMRM) model implicitly imputes missing data under MAR assumption. For variables with only one post-baseline assessment, a missing post-baseline value will be imputed by the baseline value. Patients without any post-baseline assessment will not be included in the analysis.

Histologic endpoint at the week 48 or early termination visit after at least 24 weeks of exposure to study treatment will be used for secondary efficacy analysis regardless of intercurrent event(s). Patients will otherwise be counted as non-responders.

Missing body weight at week 48 will be imputed by the last non-missing assessment, including baseline value, before week 48 (LOCF).

2.8 Safety analyses

All safety analyses will be on SAF and presented by study treatment actually received (Section 2.2).

If not clarified otherwise, safety summaries will include only data from the on-treatment period except that baseline data may also be summarized where appropriate (e.g. change from baseline summaries). The on-treatment period lasts from the date of first administration of study treatment to 30 days after the date of the last dose date.

2.8.1 Adverse events (AEs)

All reported AEs will be coded based on the latest MedDRA version practically possible before database lock and clarified in footnotes of reports.

The number (%) of patients with treatment emergent adverse events (TEAE, defined as events that started after the first dose of study treatment or events that were present prior to start of study treatment but increased in severity during on-treatment period) will be summarized by study treatment and in the following ways:

- by primary system organ class (SOC) and preferred term (PT).
- by primary SOC, PT and maximum severity.
- by Standardized MedDRA Query (SMQ) and PT.
- by SOC and PT for study treatment related adverse events,
- by SOC and PT for serious adverse events (SAE),
- by SOC and PT for TEAE leading to study treatment discontinuation.

A patient with multiple adverse events of same SOC or PT will only be counted once towards the specific SOC or PT and according to the greatest severity reported.

Clinical Trial Safety Disclosure

Regulatory-required safety disclosure tables will be produced with the final CSR.

For the legal requirements of ClinicalTrials.gov and EudraCT, two required tables on <on-treatment/treatment emergent> adverse events which are not serious adverse events with an incidence greater than 5% and on <on-treatment/treatment emergent> serious adverse events and SAE suspected to be related to study treatment will be provided by system organ class and preferred term on the safety set and for the overall study period.

If for a same patient, several consecutive AEs (irrespective of study treatment causality, seriousness and severity) occurred with the same SOC and PT:

- a single occurrence will be counted if there is ≤ 1 day gap between the end date of the preceding AE and the start date of the consecutive AE
- more than one occurrence will be counted if there is > 1 day gap between the end date of the preceding AE and the start date of the consecutive AE

For occurrence, the presence of at least one SAE / SAE suspected to be related to study treatment / non SAE has to be checked in a block e.g., among AE's in a ≤ 1 day gap block, if at least one SAE is occurring, then one occurrence is calculated for that SAE.

The number of deaths resulting from SAEs suspected to be related to study treatment and SAEs irrespective of study treatment relationship will be provided by SOC and PT.

2.8.1.1 Adverse events of special interest / grouping of AEs

Number (%) of patients with treatment emergent pruritus, diarrhea, hypoglycemia and pancreatitis) will be summarized separately for each event(in total and by the PTs contributing to it) by study treatment. Kaplan-Meier curves for the time (days) from baseline to patient's first incidence of pruritus, diarrhea and hypoglycemia by study treatment will also be provided

separately for each of the three AEs. [redacted]

Number (%) of patients who experienced following safety events will be presented by study treatment separately.

- liver safety events identified according to protocol section 16.2
- renal safety events identified according to protocol section 16.3

2.8.2 Deaths

Patients who died any time during the study will be listed with the date (study day) and reason of death as well as the preferred term, verbatim term, start date (study day) and action taken of SAEs leading to such deaths.

2.8.3 Laboratory data

Summary statistics of values and changes from baseline will be provided by visit (AV2) and by study treatment for laboratory tests listed in Table 2-6. Change from baseline will be summarized for patients with both baseline and post baseline values.

Table 2-6 Laboratory safety endpoints

laboratory test group	Parameters
Hematology	Hemoglobin (Hb), Hematocrit (Hct), Red blood cells (RBC), Mean corpuscular volume (MCV), Platelets, White blood cells (WBC), absolute WBC differential counts ((Neutrophils, Lymphocytes, Monocytes, Eosinophils, Basophils separately)
General Clinical Chemistry	Sodium, potassium, chloride, bicarbonate, calcium, magnesium, phosphorus, uric acid, haptoglobin
Liver	ALT, AST, ALP, GGT, creatine kinase (CK), lactate dehydrogenase (LDH), total bilirubin, albumin, total protein
Renal	BUN, creatinine, estimated glomerular filtration rate (eGFR)
[redacted]	
Urinalysis	Albumin-creatinine ratio (ACR), Protein-creatinine ratio (PCR)
Coagulation	INR, APTT, Prothrombin Time (PT) and Thrombin Time (TT)
Lipids and vascular risk	Total cholesterol, HDL-C, LDL-C, VLDL-C, triglycerides, free glycerol and free fatty acids, ApoAI, high sensitivity C-reactive protein (hsCRP)

Shift tables for normal / abnormal (low and high based on the low/normal/high classifications from central lab) from baseline to the most extreme post-baseline value will be provided.

Number (%) of patients with post-treatment changes in laboratory criteria related to hepatotoxicity or nephrotoxicity criteria (Section 5.3.1) will also be summarized by criteria and by study treatment. In addition, patients meeting a criterion for hepatotoxicity, nephrotoxicity or other notable abnormalities defined in Section 5.3.1 will be listed, including the criteria met as well as the laboratory test name, date (study day) of lab sampling, results and units for that test and related lab tests throughout the study.

For summaries of shift, hepatotoxicity/nephrotoxicity events or other notable laboratory abnormalities outlined above, all post-treatment data will be used regardless of last dose date.

2.8.4 Other safety data

2.8.4.1 ECG and cardiac imaging data

The PR, QRS, QT, QTcF, and RR intervals will be summarized by visit and by study treatment.

Number (%) of patients shifting from baseline in the overall normal or abnormal ECG abnormality will be presented by visit and by study treatment.

The Fridericia QT correction formula (QTcF) will be used for clinical decisions and for analyses. Notable QTcF values and changes from baseline will be summarized as categorical variable by visit and by study treatment, where a notable value is defined as a QTcF interval of greater than 450 ms and the categories used for the change (increase) in QTcF are: ≤ 30 ms, > 30 to ≤ 60 ms and > 60 ms.

2.8.4.2 Vital signs

The value and change from baseline in vital sign parameters (i.e. pulse rate, systolic blood pressure and diastolic pressure) and anthropometric measurements (i.e. body weight, waist circumference and hip circumference) and derivations (BMI, waist:hip ratio) will be summarized by visit and by study treatment.

Patients with clinically notable vital signs as defined in Section 5.3.10 will be listed with the abnormality criteria met as well as the specific test name, date (study day) of measurement, results and units for that and related measurements throughout the study.

[redacted]

Descriptive summaries of values and changes from baseline in pre-dose FGF19 and C4 will be presented by visit and by study treatment. The geometric mean and coefficient of variation will also be presented in the summaries for concentration values. The time (hours) elapsed since previous dose of study treatment will also be summarized by visit and by study treatment together in the same output.

[redacted]

[redacted]

2.14 Interim analysis

Not Applicable.

3 Sample size calculation

The sample size of the study is determined based on power consideration for proving superiority of the tropifexor-licogliflozin combination therapy compared to at least one of the monotherapies.

3.1 Power considerations for primary analysis

No data is currently available for histological responses after licogliflozin or tropifexor monotherapy, neither is there any definitive theory established for complementary or antagonist interactions between the two molecules. Historical studies published so far also suggest unignorable placebo effect ranging from 8% to 18% for different populations and endpoints, which introduces additional uncertainties about the add-on effect size of the tropifexor-licogliflozin combination therapy compared to either monotherapy even under a bliss independence assumption.

Table 3-1 Placebo effects on histological responses

Histological Response definition	Time point	Histological study	NASH population with Fibrosis stages	Placebo response rate
Fibrosis improvement of 1 stage and no worsening in NASH	Month 18	REGENERATE (2019) Obeticholic acid	F1 + F2 + F3	10.6%
			F2 + F3	11.9%
	Week 52	CENTAUR (2018) Cenicriviroc	F1 + F2 + F3	10.4%
	Week 52	ARREST (2018) Aramchol	F1 + F2 + F3	17.5%
	Week 48	STELLAR3 (2019) Selonsertib	F3	13.2%
	Week 48	STELLAR4 (2019) Selonsertib	F4	12.8%
NASH resolution with no worsening in Fibrosis	Month 18	REGENERATE (2019) Obeticholic acid	F1 + F2 + F3	8%
			F2 + F3	7.9%
	Week 52	GOLDEN (2016) Elafibranor	F0+F1+F2+F3	9%

The response rates for the monotherapies are deemed less likely to exceed 30%. For the stepwise gatekeeping testing procedure adopted for primary analysis, total sample size of 210 (70:70:70) is chosen so that there is approximately 80% power for the first hypothesis test (combination vs licogliflozin) if the true add-on effect is 20% for combination therapy.

Table 3-2 Power under different assumptions of mono and combination therapy effect sizes

True response rate assumptions			Add-on effect size in 1st test	power of 1-sided hierarchical tests α=0.05	
Licogliflozin	Tropifexor	Combination		1st test	Both tests
30%	30%	38%	8%	25%	12%
23%	23%	33%	10%	38%	22%
30%	30%	44%	14%	53%	34%
30%	30%	47%	17%	67%	52%
23%	30%	41%	18%	75%	34%
20%	20%	40%	**20%**	**84%**	75%
25%	25%	45%	**20%**	**80%**	69%
30%	30%	50%	**20%**	**79%**	67%
30%	30%	51%	21%	82%	72%
23%	30%	46%	23%	89%	60%

Power calculations were performed with simulations (n=100000) implemented in R 3.3.3, where the primary analysis in each simulated study was performed using z-tests of difference in proportions without continuity correction for simplicity.

The dropout rate of the study is anticipated to be around 10%. Because early termination and missing response data will be incorporated into the primary analysis with multiple imputations, no adjustment for early dropouts from study were implemented for the sample size calculations.

3.2 Power considerations for key secondary analysis

While key secondary efficacy endpoints will be further tested in a hierarchical order if both null hypothesis in the primary analysis are rejected, no power evaluation was performed for secondary analysis given the anticipated diminishing power after two primary hypothesis tests.

4 Change to protocol specified analyses

Analysis details, as well as editorial and semantic changes where appropriate, are implemented for different sections without altering the planned methods of analysis. The following significant changes from protocol are implemented in this SAP.

4.1 For the estimand definition in Section 2.5, one intercurrent event “Early discontinuation of study treatment or withdrawal from study before week 48 visit” is added to cover the two separately listed events "early stop of study treatment before week 24" and "premature study discontinuation before week 24" in original protocol. This updated intercurrent event also include the situation of early stop of study treatment or withdrawal from study after week 24 but before week 48. Such situation was implicitly considered as an intercurrent event, addressed with imputation but not explicitly listed as an intercurrent event in study protocol. This update does not change the primary analysis planned in the study protocol.

4.2 For the intercurrent events associated with prohibited medications, any alcohol usage more than what is permitted in the protocol, initiation of other T2DM or lipid lowering medication (e.g. GLP-1 modulator) or initiation of other treatment being evaluated or approved for NASH are also included as intercurrent event. This change is in consideration of newly identified interventions that may confound study treatment effect but not listed in the current version of study protocol.

4.3 Key secondary analyses are added into the hierarchical testing procedure after the primary analysis (Section 2.6). This change enables potentially more statistical evidence accumulation without affecting the primary analysis.

4.4 Formulation of the null and alternative hypotheses in Section 2.5 and 2.6 are updated based on odds ratio instead of difference as the test statistics. This update does not change the analysis method originally planned in protocol but to keep consistent with the actual test statistics that will be used for the hypothesis tests.

4.5 P-values will not be provided if formal hypothesis test is not performed. This change is implemented to avoid potential confusions or misinterpretation of the results.

4.6 For secondary [redacted] analysis using repeated measurement ANCOVA model, inclusion of interaction between time (visit) and treatment are explicitly clarified. This change is to ensure consistent implementation of the analysis to account for the possibility that treatment effects change over time.

4.7 Additional sensitivity analysis for the primary endpoint based on CMH test is added (Section 2.5.4), given its robustness against risk of potential misassumption for the logistic regression for primary analysis. This change is implemented also in response to FDA comments.

4.8 Multiple imputations will be performed based on FCS with logistic regression instead of discriminant function (Section 2.5.3). This change is implemented for the convenience of including dichotomous predictors in the imputer’s model.

4.9 Multiple imputations will be implemented by excluding treatment from the imputer's model instead of performing MI separately within each treatment group (Section 2.5.3). This change helps to reduce the computational burden during implementation and enables more complete cases for the MI.

4.10 Deaths will be listed instead of presented in aggregated summary (Section 2.8.2). This change is implemented because limited incidence of deaths is expected during the study. In addition, a listing provides the convenience of including additional patient specific data for the evaluation of each case of death.

4.11 Demographics and baseline characteristics summary will be only for FAS (Section 2.3). This change is implemented in the spirit of lean SAP because limited, if any, difference between SAF and FAS is expected.

4.12 Listing of safety data pre- or post treatment will not be produced. This change is implemented in the spirit of lean SAP and such listings are considered not essential for the objectives of the study. To support data review and writing of narratives, informal listings can be used.

5 Appendix

5.1 Imputation rules

5.1.1 First and last dose dates of study treatment

Partially missing start or stop date of study treatment is not expected given the eCRF design.

Missing first dose date for patients with known exposure to study treatment will be resolved through data queries.

For patients with non-missing first dose date, a missing last dose date (e.g. due to lost to follow up) will be imputed with patient's first dose date or the last known exposure to study treatment according to change in dosing CRF page, whichever is later.

5.1.2 AE start and stop dates

AE Start Date Imputation:

AE start date with missing year or with missing month but non-missing day will be resolved through data queries.

Defining AE reference date as

- If AE end date (after imputation if needed) <= first dose date, AE reference date = min (informed consent date, earliest visit date of SVs)
- Else AE reference date = first dose date

and following date components

	Day	Month	Year
Partial AE Start Date		MON	YYYY
AE reference Date	REFD	REFM	REFY

incomplete AE start date will be inputted according to logic matrix

(Imputation)	MON MISSING	MON < TRTM	MON = TRTM	MON > TRTM
YYYY MISSING	**NC**	**NC**	**NC**	**NC**
YYYY < TRTY	**(D)**	**(C)**	**(C)**	**(C)**
YYYY = TRTY	**(B)**	**(C)**	**(B)**	**(A)**
YYYY > TRTY	**(E)**	**(A)**	**(A)**	**(A)**

where

NC	No convention - require data query
(A)	01-MON-YYYY
(B)	AE reference date
(C)	Last day of month MON-YYYY
(D)	Last day of year YYYY
(E)	01-JAN-YYYY

The algorithms will be reflected in the Variable Source Derivation column as #IMPUTAEV (*event*) where *event* is the incomplete start date of the adverse event.

AE End Date Imputation:

If all AE end date components are missing after data queries, the AE is considered ongoing for analysis purpose.

If AE end month is missing, the AE end date is imputed as the last day of AE end year or patient's study completion/discontinuation visit date, whichever is earlier.

If AE day is missing, then AE end date is imputed as the last day of the month or patient's study completion/discontinuation visit date, whichever is earlier.

If imputed AE end date is less than the AE start date, use the AE start date as the imputed AE end date.

Impute Date Flag:

If year of the imputed date ≠ YYYY then date flag = Y

else if month of the imputed date ≠ MON then date flag = M

else if day of the imputed date ≠ day of original date then date_flag =

D else date flag = null

5.1.3 Start and stop dates of prior/concomitant medications or non-drug therapies / procedures

The imputation for incomplete start or stop date of medications or non-drug therapies / procedures patients received in addition to study treatment during the study will follow the same algorithms for imputations of incomplete AE start/stop date.

5.1.4 Other imputations

Imputation of missing primary or secondary endpoints are included in corresponding sections 2.5, 2.6 and 2.7 for the corresponding analyses. There will be no imputation for other analysis in this SAP.

5.2 AEs coding/grading

AEs are coded using the latest version of MedDRA dictionary practically implementable before study database lock.

5.3 Derivations

5.3.1 Laboratory parameters

Laboratory criteria of potential hepatotoxicity for liver safety monitoring in protocol section 10.2.1 (not including additional non-lab criteria):

- ALT or AST > 5 × ULN

- ALP > 2 × ULN
- TBL > 2 × ULN
- ALT or AST > 3 × ULN and INR > 1.5
- ALT or AST > 3 × ULN and TBL > 2 × ULN AND ALP to ≤ 2 × ULN (Potential Hy's Law cases)
- ALT or AST > 3 × ULN

Laboratory criteria of potential nephrotoxicity for renal safety monitoring in protocol section 10.2.2:

- Two or more consecutive serum creatinine increase >=25% compared with baseline
- serum creatinine increase >=50% compared with baseline
- eGFR < 60 mL/(min*1.73m^2)
- Two or more consecutive albumin- or protein-creatinine ratio increase ≥ 2-fold
- Two or more consecutive albumin-creatinine ratio (ACR) ≥ 30 mg/g;
- Two or more consecutive protein-creatinine ratio (PCR) ≥ 150 mg/g
- Dipstick glycosuria ≥ 1+ for patients with negative reading at baseline
- Dipstick hematuria ≥ 1+ not due to trauma for patients with negative reading at baseline

Criteria for other notable laboratory abnormality:

Table 5-1 Notable criteria for other laboratory parameters

Parameter	Threshold value	Unit
Albumin	<32	g/L
Hemoglobin	<70	g/L
Hemoglobin	>200	g/L
White blood cell count	<2.0	10^9/L
White blood cell count	>35.0	10^9/L
Platelets	<50	10^9/L
Platelets	>1000	10^9/L
INR	>4.0	
PT	>40.0	sec
APTT	>80.0	sec
Sodium	<120	mmol/L
Sodium	>160	mmol/L
Potassium	<3.0	mmol/L
Potassium	>6.0	mmol/L
Glucose	<2.2	mmol/L
Glucose	>27.8	mmol/L
Calcium	<1.50	mmol/L
Calcium	>3.00	mmol/L
Phosphate	<0.29	mmol/L
Creatinine	>177	µmol/L

Calculated eGFR	<60	mL/min

[redacted]

5.3.7 T2DM status

The T2DM status at baseline is determined as follows: T2DM = Yes if:

- Type 2 diabetes mellitus is reported in Medical History OR
- Baseline fasting plasma glucose (FPG) >= 7.0 mmol/L (126 mg/dL)

Otherwise T2DM = No.

5.3.8 Total NAS score

NAS score is calculated as the sum of scores of steatosis (0-3), lobular inflammation (0-3) and hepatocyte ballooning (0-2).

[redacted]

5.3.10 Vital signs

Table 5-2 Notable abnormalities in vital signs

Vital signs		Notable abnormalities	
		Absolute	Relative to baseline
Pulse rate (beats/min)		> 130	≥ 120 and increase from baseline ≥ 15
		< 40	≤ 50 and decrease from baseline ≥ 15
Blood pressure (mmHg)	Systolic	> 200	≥ 180 and increase from baseline ≥ 20
		<75	≤ 90 and decrease from baseline ≥ 20
	Diastolic	> 115	≥ 105 and increase from baseline ≥ 15
		< 40	≤ 50 and decrease from baseline ≥ 15

5.3.11 Definitions for biopsy based endpoints

Assessment of fibrosis staging and NAFLD Activity Scores (NAS) including specifically steatosis, lobular inflammation and hepatocyte ballooning will be performed by pathological Central Reader(s).

5.3.11.1 Improvement of fibrosis compared to baseline

The determination of fibrosis improvement will be based on NASH CRN staging. Only main stages (0, 1, 2, 3, 4) will be considered. For example, a change from 1c to 1b or 1a will not be counted as a one point change.

Table 5-3 Fibrosis stages and possible outcomes

Baseline	Week 48	1 point improvement	2 point improvement
2	0	Yes	Yes
2	1	Yes	No
2	≥2	No	No
3	≤1	Yes	Yes
3	2	Yes	No
3	≥3	No	No

5.3.11.2 Worsening of NASH

Worsening of NASH in primary estimand is defined as meeting one of the two criteria below.

- Any increase in lobular inflammation score

- Any increase in hepatocyte ballooning score

[redacted]

5.3.11.3 Resolution of NASH

Resolution of NASH in primary estimand is defined as meeting both criteria below.

- Lobular inflammation score ≤ 1
- Hepatocyte ballooning = 0

[redacted]

5.4 Statistical models

5.4.1 Logistic regression model for primary and key secondary analysis

Hierarchical testing procedure and the hypothesis tests at each step for primary and key secondary analyses are defined in Section 2.5 and Section 2.6.

For binary dependent variable and considering the link of NASH with many metabolic risk factors, a logistic regression model is used to allow more flexibility in adjusting for both categorical and continuous explanatory variables.

The efficacy of LIK066 for NASH is postulated due to its ability to improve obesity and diabetes. Frequently regarded as two manifestations of metabolic syndrome, T2DM and NAFLD commonly exist together with, for example, insulin resistance as one of the key pathological factors. The randomization of this study is stratified by patient's T2DM status at baseline. Links between obesity and histologic NASH were also well known (Lassailly et al 2015, Vilar-Gomez et al 2015).

Table 5-4 Association of weight loss with NASH resolution or liver fibrosis improvement

Study	Intervention	population	Histologic endpoint	effect
Vilar-Gomez et al 2015	Weight loss by life style change	Obese w. NASH	Resolution of NASH	p <0.01[a]
			Fibrosis status	p<0.00[a]
Lassailly et al 2015	Bariatric surgery for weight loss	Obese w. NASH	NASH resolution	85%
			Fibrosis improvement	p<0.0001[b]
			Ballooning	p<0.0001[b]
			Lobular inflammation	p<0.0001[b]

a: p-value based on Cochran-Mantel-Haenszel χ^2 trend test between weight loss categories and the histologic endpoint stratified by age older than 55 years, sex, type 2 diabetes, BMI>=35

b: p-value based on Wilcoxon signed rank test for paired data before and 1 year after bariatric surgery

Hence BMI and T2DM status are anticipated to be prognostic of histologic changes, with or without therapeutic intervention, and will be included in the logistic regression model as confounding factors.

LJN452 is anticipated to improve liver fibrosis, with or without NASH resolution, based on previous studies of other FXR modulators (Neuschwander Tetri et al 2015, Younossi 2019). The baseline stage of fibrosis is a plausible prognostic factor for histologic response, especially for the fibrosis improvement itself, and included also as a factor in the logistic regression model.

Table 5-5 FXR modulation for NASH resolution or fibrosis improvement

			Patients with histologic improvement in NASH over control		
Study	Intervention (Target/MOA)	Population	Endpoint	Diff. over control	Statistical significance
FLINT (Neuschwander et. Al. 2015)	Obeticolic acid (FXR / NAFLD)	Borderline or definite NASH	Fibrosis improvement	35% vs. 19%	p=0.004[a]
			Resolution of definite NASH	22% vs. 13%	p=0.08[a]
			Hepatocellular ballooning improvement	46% vs. 31%	p=0.03 [a]
			Lobular inflammation improvement	53% vs. 35%	p=0.006 [a]
			Portal inflammation improvement	12% vs. 13%	p=0.90 [a]
REGENERATE (Youinossi 2019)	OCA 25mg (FXR/ NASH)	NASH w stage 2/3 fibrosis	Fibrosis improvement w/o NASH worsening	21% vs. 10.6%	p=0.0002[b]
			NASH resolution w/o fibrosis worsening	11.7% vs. 8.0%	p=0.1268[b]

a: p-value based on Cochran-Mantel-Haenszel χ^2 trend test stratified by clinic and diabetes status

b: no details on calculation of p-value in the press release

The logistic regression model assumes patient's week 48 histological response status y ∈(0, 1) ~ binomial (1, p_{jklm}), where p_{jklm} is the probability of histologic response and related to the collection of factors/covariates $\boldsymbol{X}$ corresponding to patient's study treatments *j*, baseline fibrosis stage *k*, baseline T2DM status *l* and baseline BMI *m* through logit link function (Nelder and Wedderburn 1972)

For patients sampled in this study, it follows that $\log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X_i}$ for i=(1, 2, ..., n), where n is the sample size of the study and $\boldsymbol{\beta}$ is the vector of coefficients to be estimated for the effects of independent variables.

Under the aforementioned binomial distribution, the probability function can be derived as

$$f(y_i) = {p_i}^{y_i}(1-p_i)^{1-y_i}$$

with log likelihood function

$$l = \sum_{i=1}^{n} (y_i \log(p_i) + (1 - y_i)\log(1 - p_i))$$

where $p_i = \frac{\exp(\boldsymbol{\beta^T X_i})}{1+\exp(\boldsymbol{\beta^T X_i})}$

The basic maximum likelihood estimators (MLE)$\widehat{\boldsymbol{\beta}} \sim N(\beta, I(\beta)^{-1})$, where $I(\beta)$ is the Fisher's information matrix, for such logistic regression model is known to suffer from small sample bias. Given the uncertainties on benefits of LJN452 and LIK066 over placebo effects (Table 3-1), the penalized likelihood method proposed by Firth (1993) will be adopted to reduce potential small-sample bias and to consistently produce finite MLEs in case of complete or quasi-complete separations during the calculations.

Over-dispersion associated with the estimation of variances for binomial data will be corrected using the Williams method (1982).

With randomization and blinding, the possibility of collinearity between study treatment and other independent variables can be reasonably ignored so inflation on standard errors of the coefficients for treatment effect is not expected in the analysis model despite potential collinearity among other cofactors.

5.4.1.1 Interactions among independent variables and subgroup analysis

It is possible that the different mechanisms of actions with LJN542 and/or LIK066 may lead to differential effects on the components of the primary endpoints, which is potentially an advantage achieved with combination therapy. If such differential effects exist, however, it may also lead to interactions between covariates and main treatment effect, which is not included in the primary analysis model.

The statistical inference of primary analysis is on the overall patient population targeted by this study, notwithstanding the possibility of such granular modulation of the treatment effect. Nevertheless, better understanding on what limits or enhances the add-on effect of the combination therapy, and for whom or under what circumstances it exists, could inform decisions on future development of the combination therapy.

Considering the sample size of the study and unknown treatment effects, adding additional interaction effect in the logistic model could increase the risk of small sample (small number of events/non-events within class) bias and failures to converge. Hence subgroup analysis as outlined in Section 2.13.1 will be performed. [redacted]

5.4.2 Multiple imputation for missing primary and key secondary analysis

MI method for missing data allows the flexibility of including additional auxiliary variables in the imputer model and accounts for the uncertainty on real value with multiple inputted values.

For the binary dependent variable, a Fully Conditional Specification (FCS) logistic regression method is chosen for the imputer model considering the convenience of including dichotomous predictors. Baseline fibrosis stage (2 vs. 3), baseline T2DM (No/Yes), baseline BMI (as

categorical predictor) and decrease of ALT at week 24 (<= -17 U/L vs. >-17 U/L) will be included as predictors for the imputer's model.

The exclusion of treatment from the imputer model ensures a more conservative estimate of treatment effect in that it would bias only toward null (for superiority test) if a relationship between treatment and response does exist (Schafer 1997). The inclusion of decrease in ALT at week 24 (<= -17 U/L vs. >-17 U/L) into the imputer's model is based on findings by Loomba et. al. (2019) and considering the balance between general validity of analysis model and some loss in precision in case a trivial factor is additionally added in imputer's model (Rubin 1996).

If a reduced logistic regression model has to be adopted (Section 2.5.2), the reduced list of predictors in primary analysis model will also be used for the imputer model, except for the treatment which will still be excluded.

The imputer's model is first fitted using observations (n_0) with non-missing data

$$\text{logit}\,(p) = \boldsymbol{\beta}^{\mathbf{T}}\mathbf{X}$$

where

p is the true probability of response

X is the n_0 *by p* matrix of covariates

β is the vector of coefficients to be estimated as the first step of MI.

Based on the estimated regression coefficients $\widehat{\boldsymbol{\beta}}$ and associated covariance matrix, posterior predictive distribution for **β** is built and a new set of parameters $\boldsymbol{\beta}_{\#}$ is randomly sampled from it.

For *i*th patient with missing data, expected probability q_i =Pr (response) is calculated as

$$q_i = \frac{\exp(\boldsymbol{\beta}_{\#}^{\mathbf{T}}\boldsymbol{X_i})}{1+\exp(\boldsymbol{\beta}_{\#}^{\mathbf{T}}\boldsymbol{X_i})}$$

and this patient's response status will be imputed as Yes if q_i is larger than a random number sampled from uniform distribution between 0 and 1.

These imputation steps can be implemented using SAS MI procedure. A total of >=50 iterations of imputations will be executed to create 50 imputed complete data sets.

The logistic regression model for primary or key secondary analysis is then executed separately for these data sets with results (parameter estimates) stored in an output data set.

The SAS MIANALYZE procedure reads in the parameter estimates, standard errors or covariance matrix in the output data set generated by the logistic regression analyses, and produces estimates and 95% confidence limits to back calculate the odds ratios (95% CI). P-values will be presented if formal hypothesis test is to be performed according to hierarchical testing procedure outlined in Sections 2.5 and 2.6.

5.4.3 Repeated measurement ANCOVA (MMRM) for secondary [redacted] analyses

Suppose the continuous dependent variable vector $y_i = (y_{i1}, \ldots, y_{im_i})'$ of a patient $i = 1,...,n$ observed at different time points (visits) $j=1, ..., m$ follows a m-dimensional multivariate normal distribution with some non-trivial off-diagonal elements in its $m_i \times m_i$ variance - covariance matrix.

The MMRM is chosen because it can explicitly account for such correlation and non-constant variability in data with a schematic model

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Z\gamma} + \boldsymbol{\varepsilon}$$

where

X ~ covariates of fixed effects (including treatment effect of interest) with coefficients vector $\boldsymbol{\beta}$

Z ~ random effects with coefficients $\boldsymbol{\gamma}$ *normally distributed with expected value* $\mathbf{0}$ *and covariance matrix (**G**)*

$\boldsymbol{\varepsilon}$ *~ is random error not accounted for in the model with expected value* $\mathbf{0}$ *and a covariance matrix (**R**)*

It follows that the variability of Y can be structured as V=ZGZ`+R for parameterization.

The parameterization of the covariance matrix structures for repeated measurements could be pre-specified based on trial design and assumptions. For analysis in this study, an unstructured covariance matrix is assumed. In case of failure to converge that cannot be resolved with other approaches, compound symmetry structure may be used if it resolves the issue.

Both covariance parameters and the fixed/random effects (including 95% confidence limits) can then be estimated based on likelihood methods, such as the restricted maximum likelihood (REML) method which accommodates the situation of some data missing at random (Rubin 1976; Little 1995) and will be adopted in this study.

The analysis can be implemented in the SAS MIXED procedure with TYPE=UN and subject=<patient ID> in the REPEATED statement and DDFM=KR in model statement. The REML is the default method.

Model assumptions will be checked based on diagnostic plots and influence diagnostics if necessary. The possible impact of questionable models will be discussed in CSR if needed.

5.4.4 CMH method for sensitivity analysis

The CMH analysis will be performed separately for combination vs. LJN452 and for combination vs. LIK066. Each CMH analysis is stratified by three binary predictors, baseline fibrosis stage (<=2 vs. >2), T2DM (No/Yes) and BMI (>= 35 vs. < 35) added to TABLES statement before treatment and response status in the SAS FREQ procedure.

The stratification factors subgroup data into a series of i=8 separate 2x2 contingency tables

Primary endpoint	Combination Therapy	Mono-therapy	Total
Responder	a_i	b_i	$t_{1.i}$

Non-responder	c_i	d_i	$t_{2.i}$
Total	$t_{.1i}$	$t_{.2i}$	T_i

A common odds ratio (OR) between combination and monotherapy in the probability of response can be estimated as proposed by Mantel and Haenszel (1959)

$$\widehat{OR} = \frac{\sum_{i=1}^{8} \frac{a_i d_i}{T_i}}{\sum_{i=1}^{8} \frac{c_i b_i}{T_i}}$$

and asymptotic 95% confidence limits for the common odds ratio can be calculated based on estimated variance for the odds ratio on log scale (Robins et. al. 1986).

The exact test for the common odds ratio = 1 will be performed using the EXACT statement with the COMOR option in the SAS FREQ procedure to report a p-value.

5.5 Rule of exclusion criteria of analysis sets

Table 5-6 Scenarios causing subjects to be excluded

Scenarios	Exclusion from
Patient was rescreened but did not sign a new ICF	All analysis sets
ICH-GCP non-compliance of study site with impact on data quality	All analysis sets
No drug taken after randomization	SAF
ICF not signed	All analysis sets
Mis-randomized and received no study treatment	FAS

5.6 Other statistical aspects

5.6.1 Risk difference and Wald asymptotic 100*(1-α)% CI

For n patients each at risk to experience a certain event with probability π, the crude incidence is estimated as p=x/n, where x is the number of patients with the event.

For two proportions $p_1 = x_1/n_1$ and $p_2 = x_2/n_2$, the risk difference $d = p_1 - p_0$ has asymptotic standard error $se(d) = \text{sqrt}\{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2\}$. The Wald asymptotic 100(1-α)% confidence limits for the difference are $d \pm se(d) * z_{\alpha/2}$. Wald asymptotic 95% CI for difference can be calculated using SAS procedure PROC FREQ and option RISKDIFF in the TABLES statement, specifying the RISKDIFF option also in the EXACT statement in case exact CI is needed for rare events.

6 References

Robins, J. M., Breslow, N., and Greenland, S. (1986). "Estimators of the Mantel-Haenszel Variance Consistent in Both Sparse Data and Large-Strata Limiting Models." *Biometrics* 42:311–323.

Mantel, N., and Haenszel, W. (1959). "Statistical Aspects of Analysis of Data from Retrospective Studies of Disease." *Journal of the National Cancer Institute* 22:719–748.

Rubin, D. B. (1976). "Inference and Missing Data." *Biometrika* 63:581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Rubin, D. B. (1996). "Multiple Imputation after 18+ Years." *Journal of the American Statistical Association* 91:473–489.

Nelder, J. A. and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society, Series A*, 135, 370–384.

Firth, D. (1993), "Bias Reduction of Maximum Likelihood Estimates," *Biometrika*, 80, 27–38.

Williams, D. A. (1982), "Extra-binomial Variation in Logistic Linear Models," *Applied Statistics*, 31, 144–148.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall.

Little, R. J. A. (1995). "Modeling the Drop-Out Mechanism in Repeated-Measures Studies." *Journal of the American Statistical Association* 90:1112–1121.

Bretz, F., Maurer, W., Brannath, W. and Posch, M. (2009). "A graphical approach to sequentially rejective multiple test procedures." *Statistics in Medicine* 28, 586–604.

Vilar-Gomez E, Martinez-Perez Y, Calzadilla-Bertot L, et al (2015) "Weight Loss Through

Lifestyle Modification Significantly Reduces Features of Nonalcoholic Steatohepatitis," *Gaestroenterology* 367-78.

Lassailly G, Caiazzo R, Buob D, Pigeyre M, Verkindt H, Labreuche J, et al. (2015) "Bariatric surgery reduces features of nonalcoholic steatohepatitis in morbidly obese patients," *Gastroenterol* 149:379–88.

Neuschwander-Tetri BA, Loomba R, Sanyal AJ, et al (2015) "Farnesoid X nuclear receptor

ligand obeticholic acid for non-cirrhotic, non-alcoholic steatohepatitis (FLINT): a multicentre, randomised, placebo-controlled trial," *Lancet* 956-65.

Youinossi Z (2019) "Positive Results from REGENERATE: A Phase 3 International, Randomized, Placebo-Controlled Study Evaluating Obeticholic Acid Treatment for NASH," Abstract Gs-06, EASL

Loomba R, Sanyal AJ, Kowdley KV, et al. (2019) "Factors Associated With Histologic Response in Adult Patients With Nonalcoholic Steatohepatitis," *Gastroenterology* 156:88-95.e5.