Early Health Technology Assessment during Non-alcoholic Steatohepatitis Drug Development: A Two-Round, Cross-Country, Multicriteria Decision Analysis

APPENDIX

METHODS

Phase 'a' - Problem Structuring: Clinical Practice and Scope

Long-term cohort studies have found the severity of liver fibrosis at baseline was the best predictor of liver-related and overall mortality in patients with NAFLD [56, 57]. Out of liver-related complications, it has been suggested that NAFLD may promote cardiovascular disease and extra-hepatic malignancy [58-60]. While liver biopsy remains the gold standard for the diagnosis of NASH, numerous biomarkers and imaging methods have been developed for patient screening and monitoring [61].

Treatment of NASH is challenging, as progression from steatosis to NASH and fibrosis is likely a multi-factorial process that is difficult to predict, involving varied molecular pathways that may dominate in different patient subsets, including insulin resistance, lipotoxicity, pro-inflammatory cytokine release from adipose, impaired lipid and cholesterol metabolism, gut microbial dysbiosis [62] and genetic background. Out of lifestyle changes, there are currently no approved pharmacological therapies for NASH and effective treatments are eagerly awaited. This implies that NAFLD is not only focused on the liver alone, it is a multifactorial disease [63, 64].

Phase 'b' - Model Building: The Advance Value Tree (AVT) Adaptation, Alternative Treatments and Evidence, Attribute Ranges and References

In terms of evidence considered, the main clinical evidence on the therapeutic and safety performance of the three late development compounds in Phase 3, was sourced from three multi-centre, double-blinded, randomised, placebo-controlled Phase 2 trials that were publicly available, whose names will not be disclosed to maintain the anonymity of the exercise. Evidence from these three trials was used to assess the performance of the three respective compounds, assuming comparable patient populations. Although the definitions for some outcomes of interest across the three trials were not identical, they were considered comparable for the purpose of the exercise. Additional non-therapeutic evidence relating to safety and innovation value dimensions was sourced from the summaries of product characteristics (SmPCs) available through the EMA European Public Assessment Reports (EPAR) (for two compounds which had already received product Marketing Authorisation (MA) for the treatment of two other indications), Anatomical Therapeutic Chemical (ATC) classification system indexes through the portal of the WHO Collaborating Centre for Drug Statistics Methodology [65], and ClinicalTrials.gov listings [66].

The lowest and highest placebo performance levels as evident across the different clinical trials of interest were also shown to the Decision Conference (DC) participants in order to provide an insight of the treatment effect versus placebo control. In France and Germany, some of the measurement units used were not the same as the ones reported in the clinical trials (especially for metabolic factors), which required us to make some conversions on the spot with the assistance of the relevant experts.

Ideally an indirect treatment comparison through a network meta-analysis would be conducted to synthesise the available clinical evidence from all the relevant clinical trials and derive relative treatment effects for the outcomes of interest (e.g. taking the form of mean

differences for continuous outcomes, and odds ratios or risk ratios for binary outcomes, in addition to their 95% confidence intervals). Such relative effects could then be translated to an absolute scale using an estimate of the absolute effect for a suitably selected baseline treatment (to make value trade-offs in a meaningful way) which could eventually be used to assess the performance of the selected compounds. However, given the small number of clinical trials available and the existence of only a single common (placebo) arm between each intervention in the network, such estimates would be highly uncertain. Due to the lack of evidence on several outcomes for some of the study arms, consistent with the early phase of the trials, and the complex meaning of relative treatment effects which could make the elicitation of value trade-offs between outcomes cumbersome, it was decided to use directly the un-synthesised evidence on the performance of the different compounds from their respective clinical trials.

Phase 'c' - Model Assessment and Phase 'd' – Model Appraisal: Decision Conferences and MCDA Technique

Decision Conferences could be defined as "a gathering of key players who wish to resolve important issues facing their organisation, assisted by an impartial facilitator who is expert in decision analysis, using a model of relevant data and judgements created on-the-spot to assist the group in thinking more clearly about the issues" [37] [67]. Typical stages of DCs include (i) exploring the issues, (ii) structuring and building the model, (iii) exploring the model and (iv) agreeing on the way forward. The first two stages were largely informed by preparatory work conducted prior to the actual DCs, involving extensive literature reviews and clinical expert consultation as part of model building phase.

Background material introducing the scope of the exercise in greater detail was sent to the DC participants one week prior to the actual DC. On the day of each DC, the aims and the

scope of the exercise together with a description of the MCDA methodology were presented, followed by a clinical overview of NASH from a hepatology specialist (MT, VR, AC, AOB, LS, IS – co-authors of the study), including epidemiology, disease progression and its burden.

The preliminary NASH value tree was then presented to the participants and revised cluster by cluster in real time through a facilitated open discussion [37, 68]. The various criteria were explained, followed by a group discussion relating to their relevance and completeness in the context of the exercise. For this purpose, the respective "lower" (x_l) and "higher" (x_h) reference levels where considered as the attribute range of interest to reflect the performance of the alternative compounds. As a result of this iterative process, some of the criteria could either be removed, because they were perceived as irrelevant or non-fundamental (e.g. due to non-meaningful clinical difference in performance), or be added, because they were deemed to be missing from the preliminary NASH value tree.

Subsequently, value functions were elicited for the different criteria by linking each attribute range with a value scale, and criteria weights were elicited for each attribute by considering their relative importance. Finally, the emerging overall Weighted Preference Value (WPV) scores of the options were analysed, by illustrating the comparative performance between different pairs of treatments across the model criteria. The ranking of the results was also tested for their robustness by conducting (a) one-way sensitivity analyses across a selected set of criteria for which consensus between participants could not be reached and (b) a multi-way sensitivity analysis (i.e. "robustness analysis") as part of which all value scores and/or criteria weights would be changed simultaneously.

M-MACBETH was used to elicit value functions for the different attributes, assign attribute weights through a qualitative swing weighting approach, aggregate the preference value scores and weights using an additive aggregation to derive overall WPV scores, and conduct sensitivity analyses [69]. In addition to a consistency check for the qualitative

judgements of the group provided automatically by the software, a further consistency check was performed manually by the facilitator to ensure that an interval scale was obtained. Specifically, in order to validate the cardinality of the scale, the facilitator compared the sizes of the intervals between the suggested scores and invited participants to adjust them if necessary [70], an essential requirement for aggregation when using simple additive value models.

The robustness of the results was also tested to understand under which conditions an option is considered to be "globally" more attractive than another by using "robustness analysis" function in the M-MACBETH software to test which types of preference information given, i.e. ordinal, pre-cardinal (i.e. MACBETH) or cardinal, is required to sustain a given option as being globally better than any other.

RESULTS

Overall Compound Rankings, Value Composition and Sensitivity Analysis

The sensitivity and robustness analyses demonstrated that treatment rankings are, in general, fairly robust across the different model settings. The lowest possible weight changes required for the other compounds to rank first (above compound D), would be an decrease of 29% in the attribute "Fibrosis improvement" at the London DC1 for compound C to rank first an increase of 501% in the attribute "ALT" at the Paris DC2 for compound B to rank first; and an increase of 1249% in the attribute "Triglycerides" at the Paris DC2 for compound A to rank first.

The one-way sensitivity analysis revealed 4 main scenarios in compound top rankings resulting from individual weight changes (i.e. a change in the weight of one criterion while keeping all remaining criteria relationships unchanged): no change in the ranking of

compound D, compound B achieving the highest rank, compound A achieving the highest rank and compound C achieving the highest rank.

The first scenario (i.e. no change in the ranking of compound D) holds true for any weight changes across all attributes in the Safety Profile (SAF) cluster and across all settings with the single exception of the attribute "Nausea, Vomit, Diarrhea" at France DC2 (85% weight reduction could place compound C first). Beyond the SAF cluster, this scenario also held true for the attributes "LDL cholesterol", "Diastolic pressure", "Gamma GT" and "SF-36 Mental Comp" across all settings. Where this scenario held true, the respective weights of the relevant attributes accounted for between 29% (England DC1) and 83% (Germany DC2) of the total model weight.

The second scenario, in which compound B is ranked first, could only hold true following extremely large changes in some attribute weights. Specifically, this could result by increasing the weight of the attribute "ALT" in five of the six DCs (with the exception of England DC1) at a range between 501% (France DC2) and 6,742% (England DC2); increasing the weight of "New indications in Ph2" by 891% at England DC1; or increasing the weight of "New indications in MA" by 4,053% or 5,991% at France DC1 or Germany DC1, respectively.

The third scenario, in which compound A is ranked first, could only result following a 1,249% increase in "Triglycerides" weight at France DC2.

Changing the weights of the remaining attributes across settings could only result in compound C becoming the highest ranked option, as part of scenario 4. For these attributes, the minimum required weight changes would range between a decrease of 29% (Fibrosis improvement at England DC1) to an increase of 350% (NASH Resolution at Germany DC2), with the respective weights of the relevant attributes accounting for between 16% (Germany DC2) and 67% (England DC1) of the total model weight.

A robustness analysis was then conducted to understand if changing combinations of attribute weights (i.e. n-way) could lead to changes in rankings at smaller % changes, this time without keeping the attribute weight relationships constant. It was revealed that for all settings with the exception of England DC1 and Germany DC1, no change in weight combinations could lead to a change in the ranking between compounds D and C (first and second), as long as the swing (MACBETH) qualitative value judgements elicited by participants remained constant. At the England DC1 and Germany DC1 settings however, a change of +/-2% on all attribute weights together could lead to a result where compound C could outrank compound D by 14.3 value points (weighted), caused by favouring all criteria where compound C outperforms compound D.

Another robustness analysis was conducted by exploring changes on all value scores of attribute value functions. A change of 10% on the scores of the levels across all attributes would not provide a new ranking with the exception of the Germany DC2 were the only change in the ranking would be between the bottom half ranks (change between compound B and A). In the case of England DC2 and Germany DC1, the results proved to be even more robust with the 3 top-ranked alternatives remaining the same even if the MACBETH value judgements provided by participants across all attribute value functions were to be ignored (relating to judgements of differences in attractiveness between references to elicit value scales).

Table A1: Compound Performance across the Criteria Attributes.

Where clinical attributes are reported in two forms, these correspond to either a "percentage change" or "absolute change" from baseline, as reflected with white and grey coloured rows respectively, essentially reflecting different measuring units; although only one of the two attribute forms was used for eliciting participant preferences, the other attribute form could still be considered if requested by the participants as an aid to expressing their value judgements.

Cluster	Criteria	Attribute name	Attribute metric	Compound A	Compound B	Comp ound C	Compou nd D	Placebo lowest	Placebo highest	Least preferred	Most preferred
Therapeu tic Impact	Histologic endpoints	**NASH resolution**	% of patients experiencing resolution of NASH without worsening of fibrosis	26%	22%	39%	32%	5%	13%	22%	39%
Therapeu tic Impact	Histologic endpoints	**Fibrosis improveme nt**	% of patients experiencing fibrosis improvement	22%	35%	26%	55%	14%	19%	22%	55%
Therapeu tic Impact	Lipids	**LDL cholesterol**	Mean % change from baseline in LDL cholesterol, mmol/L	-7.4%	7.6%	-3.8%	-7.6%	-7.6%	7.1%	7.6%	-7.6%
Therapeu tic Impact	Lipids	**LDL cholesterol**	Mean absolute change from baseline in LDL cholesterol, mmol/L	-0.20	0.22	-0.10	-0.22	-0.22	0.20	0.22	-0.22
Therapeu tic Impact	Lipids	**HDL cholesterol**	Mean % change from baseline in HDL cholesterol, mmol/L	5.0%	-1.8%	6.4%	0.0%	-3.8%	2.7%	-1.8%	6.4%
Therapeu tic Impact	Lipids	**HDL cholesterol**	Mean absolute change from baseline in HDL cholesterol, mmol/L	0.06	-0.02	0.07	0.00	-0.05	0.03	-0.02	0.07

Therapeutic Impact	Metabolic factors	**HbA1c**	Mean % change from baseline in Glycated Haemoglobin A1c, mmol/mol	-1.0%	1.0%	-13.8%	0.0%	0.0%	0.9%	1.0%	-13.8%
Therapeutic Impact	Metabolic factors	**HbA1c**	Mean absolute change from baseline in Glycated Haemoglobin A1c, mmol/mol	-0.46	0.50	-5.70	0.00	0.00	0.40	0.50	-5.70
Therapeutic Impact	Metabolic factors	**Body weight**	Mean % change from baseline in Body Weight	0.0%	-2.3%	-5.2%	0.0%	-0.56%	0.00%	0.0%	-5.2%
Therapeutic Impact	Metabolic factors	**Body weight**	Mean absolute change from baseline in Body Weight	0.00	-2.30	-5.30	0.00	-0.60	0.00	0.00	-5.30
Therapeutic Impact	Metabolic factors	**Systolic blood pressure**	Mean % change from baseline in Systolic Pressure, mm Hg	0.0%	-3.0%	-3.8%	0.0%	-2.3%	-0.8%	0.0%	-3.8%
Therapeutic Impact	Metabolic factors	**Systolic blood pressure**	Mean absolute change from baseline in Systolic Pressure, mm Hg	0.0	-4.0	-5.0	0.0	-3.0	-1.0	0.0	-5.0

Therapeutic Impact	Metabolic factors	**Diastolic blood pressure**	Mean % change from baseline in Diastolic Pressure, mm Hg	0.0%	0.0%	0.8%	0.0%	0.0%	3.1%	0.8%	0.0%
Therapeutic Impact	Metabolic factors	**Diastolic blood pressure**	Mean absolute change from baseline in Diastolic Pressure, mm Hg	0.0	0.0	0.6	0.0	0.0	2.4	0.6	0.0
Therapeutic Impact	Metabolic factors	**ALT**	Mean % change from baseline in Alanine aminotransferase, U/L	-18.8%	-45.8%	-35.1%	-18.8%	-3.1%	-22.0%	-18.8%	-45.8%
Therapeutic Impact	Metabolic factors	**ALT**	Mean absolute change from baseline in alanine Aminotransferase, U/L	-12.0	-38.0	-27.0	-12.0	-2.0	-18.0	-12.0	-38.0
Therapeutic Impact	Metabolic factors	**Triglycerides**	Mean absolute change from baseline in Triglycerides, grams/L	-0.44	-0.19	-0.02	-0.19	0.18	-0.07	-0.02	-0.44
Therapeutic Impact	Metabolic factors	**GGT**	Mean absolute change from baseline in Gamma-Glutamyl Transferase, U/L	-22.0	-37.0	-33.7	-37.0	-6.0	-7.2	-22.0	-37.0

Therapeutic Impact	Metabolic factors	**HOMA-IR**	Mean absolute change from baseline in HOMA-IR, units	15.0	15.0	-1.8	15.0	4.0	0.7	15.0	-1.8
Therapeutic Impact	Quality of Life	**SF-36 Physical comp**	Mean % change from baseline, score	0.0%	0.0%	4.2%	0.0%	-2.3%	-1.3%	0.0%	4.2%
Therapeutic Impact	Quality of Life	**SF-36 Physical comp**	Mean absolute change from baseline, score	0.0	0.0	1.9	0.0	-1.0	-0.5	0.0	1.9
Therapeutic Impact	Quality of Life	**SF-36 Mental comp**	Mean % change from baseline, score	0.0%	0.0%	-5.5%	0.0%	-7.3%	2.1%	-5.5%	0.0%
Therapeutic Impact	Quality of Life	**SF-36 Mental comp**	Mean absolute change from baseline, score	0.0	0.0	-2.8	0.0	-3.3	1.0	-2.8	0.0
Safety profile	Adverse Events	**Treatment related Serious AEs**	% of patients experiencing treatment-related Serious Adverse Events	2%	4%	0%	0%	0%	4%	4%	0%
Safety profile	Adverse Events	**Overall Serious AEs**	% of patients experiencing Serious Adverse Events	16%	21%	8%	0%	8%	15%	21%	0%
Safety profile	Adverse Events	**Nausea**	% of patients experiencing Nausea	10.1%	8.5%	46.0%	0%	8%	38%	46%	0%
Safety profile	Adverse Events	**Pruritus, G1-G2**	% of patients experiencing Pruritus, G1-G2	1.1%	21.3%	0.0%	0%	2%	6%	21%	0%

Safety profile	Adverse Events	**Pruritus, G3**	% of patients experiencing Pruritus, G3	0.0%	2.1%	0.0%	0%	0%	2%	2%	0%
Safety profile	Adverse Events	**Pruritus, G2-G3**	% of patients experiencing Pruritus, G2-G3	1.1%	17.0%	0.0%	0%	2%	2%	17%	0%
Safety profile	Adverse Events	**Pruritus, any**	% of patients experiencing Pruritus	1.1%	23.4%	0.0%	0%	2%	6%	23%	0%
Safety profile	Adverse Events	**Renal AE**	% of patients experiencing Renal Adverse Events	6.7%	7.1%	7.7%	0%	0%	7.69%	8%	0%
Safety profile	Adverse Events	**Fatigue AE**	% of patients experiencing Fatigue	5.6%	0.0%	15.4%	0%	0%	19.23%	15%	0%
Safety profile	Adverse Events	**Cardio AE**	% of patients experiencing Cardiovascular Adverse Events	0.0%	9.2%	11.5%	0%	0%	7.69%	12%	0%
Safety profile	Adverse Events	**Nausea, vomit, diarrhea AE**	% of patients experiencing Nausea, Vomit or Diarrhea	19.1%	8.5%	103.8 %	4%	7.75%	69.23%	104%	0%
Safety profile	Contraindications & warnings	**Contraindications**	Existence of any Contraindications of use	None known	None	Type 1 diabetes mellitus	None known	NA	NA	T1DM	None known

Safety profile	Contraindications & warnings	**Drug-drug interactions**	Existence of any drug–drug interactions	None known	May reduce warfarin INR, increase exposure to CYP1A2 substrates, reduced efficacy by bile acid binding resins	Very low potential for PK interactions with CYP450 substrates	None known	NA	NA	warfarin, CYP1A2 substrates, bile acid binding resins	None known
Innovation level	Mechanism of Action	**ATC L4**	The technology's mechanism of action innovation type as reflected though WHO's Anatomical Therapeutic Chemical (ATC) Classification System, Level 4	1st PPAR α/δ agonist (anti-inflammatory)	1st bile-acid derivative FXR agonist (multi-modal)	1st GLP-1 agonist (anti-steatotic)	2nd FXR agonist (multi-modal) but first non-bile acid FXR-agonist	NA	NA	NA	NA
Innovation level	Spill-over effect	**New indications in Ph1**	Number of other indications investigated as part of Phase 1 clinical trials	3 (T2DM, Obesity, Dyslipidemia)	2 (Renal Impairement, Obesity)	2 (T2DM, Obesity)	0	NA	NA	0	3

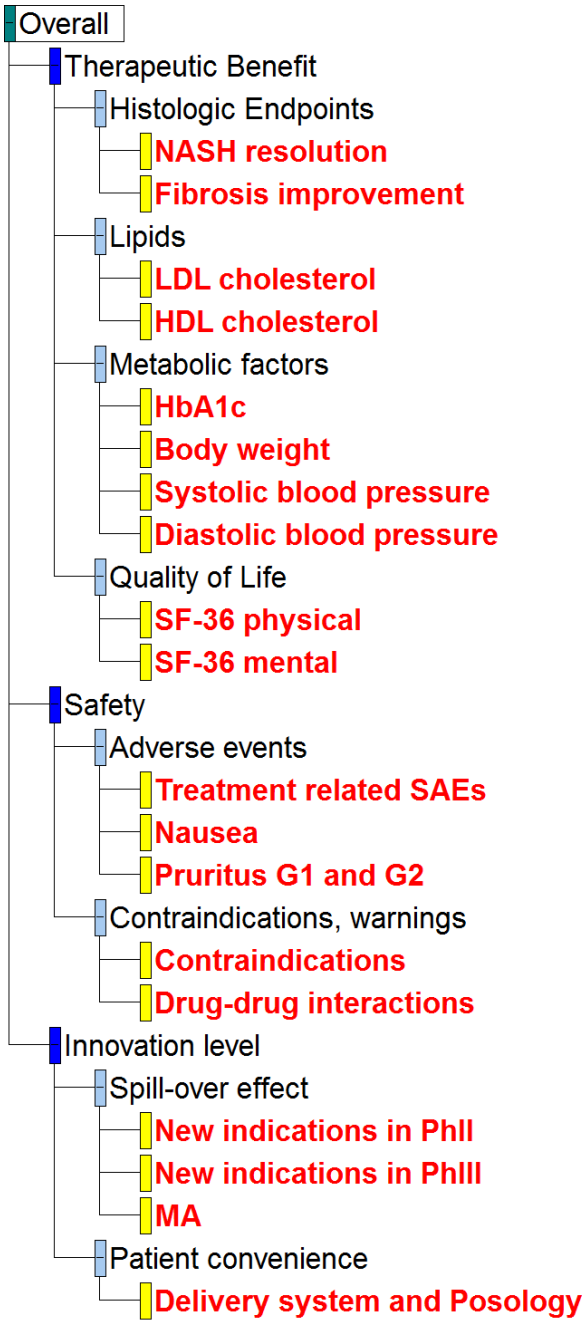
Innovation level	Spill-over effect	**New indications in Ph2**	Number of other indications investigated as part of Phase 2 clinical trials	3 (T2DM, Obesity, Dyslipidemia)	8 (Chronic Diarrhea, Familial Partial Lipodystrophy,Alcoholic Hepatitis, Obesity, Gallstones, PBC, Primary Sclerosing Cholangitis, Diabetes and Presumed NAFLD)	1 (Obesity)	2 (Primary Biliary Cirrhosis, Primary Bile Acid Diarrhea)	NA	NA	1	8
Innovation level	Spill-over effect	**New indications in Ph3**	Number of other indications investigated as part of Phase 3 clinical trials	0	0	1 (T2DM)	0	NA	NA	0	1
Innovation level	Spill-over effect	**New indications in MA**	Number of other indications having received Marketing Authorisation (MA)	0	1 (Primary Biliary Cholangitis)	0	0	NA	NA	0	1
Innovation level	Patient convenience	**Delivery system and posology**	Combination of the delivery system and the posology (i.e. dosing)	Oral once daily	Oral once daily	Subcutaneous once daily	Oral once daily	NA	NA	Subcutaneous once daily	Oral once daily

Table A2: Raw Data for Sensitivity Analysis (see online supplementary excel file).

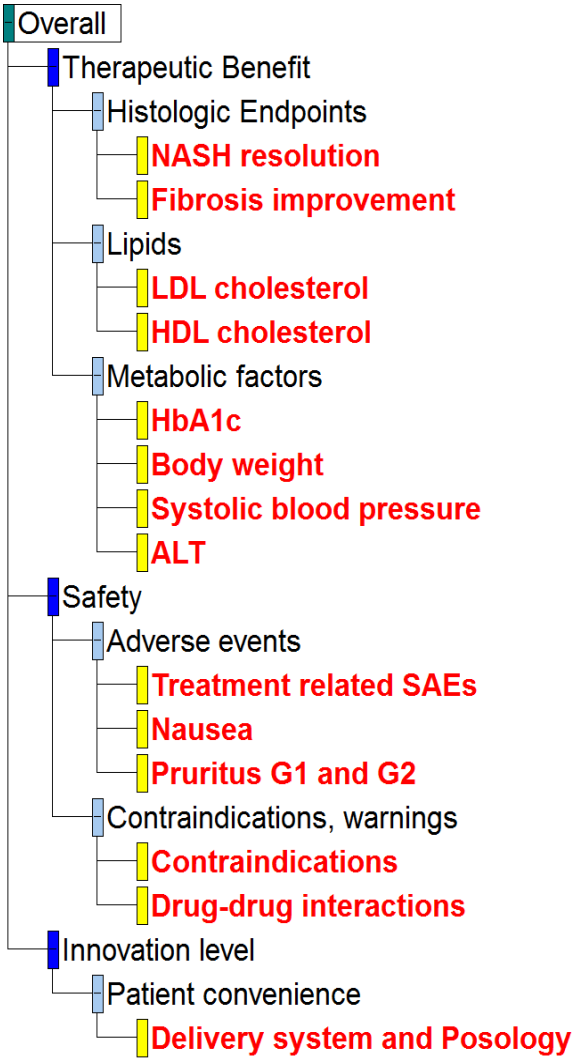
Figure A1: Final Value Trees following the Two Rounds of Decision Conferences for a) England, b) France and c) Germany (Round 1 and Round 2, respectively)

a)

Round 1

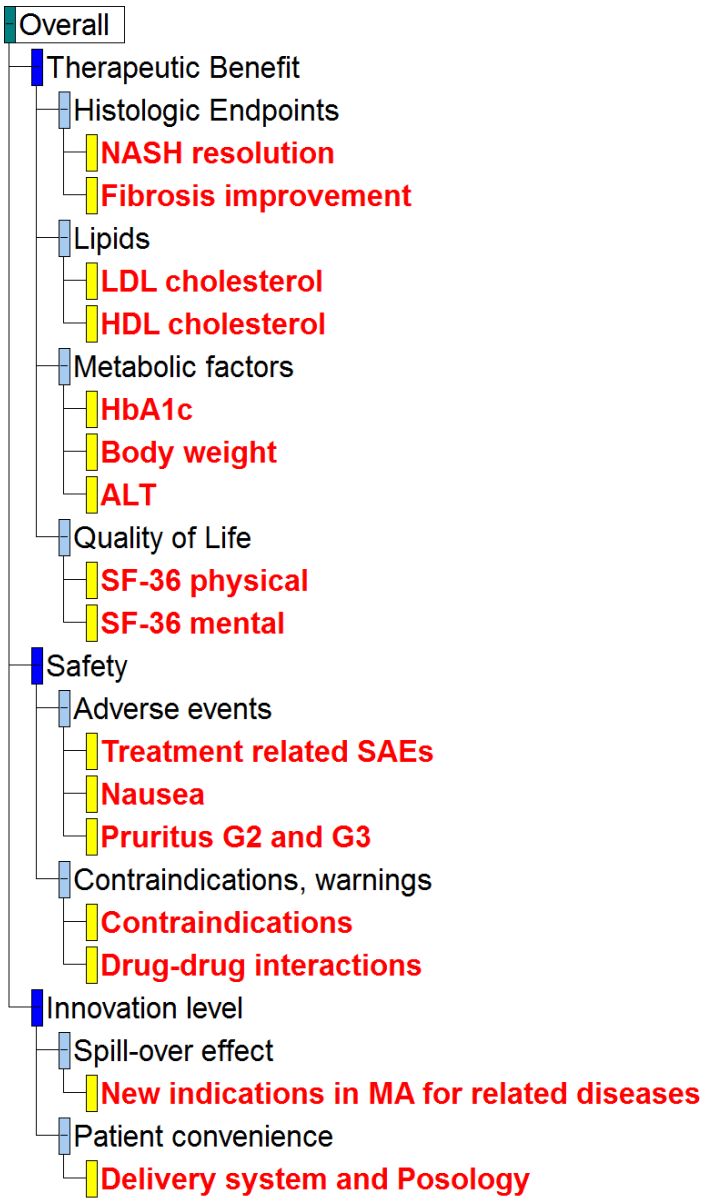


Round 2

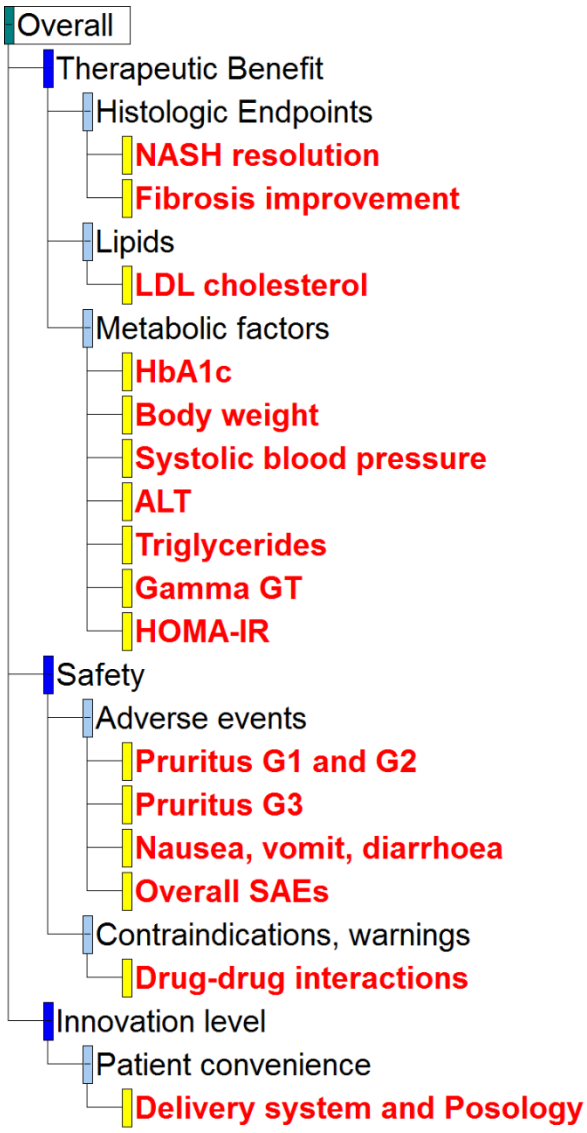


b)

Round 1

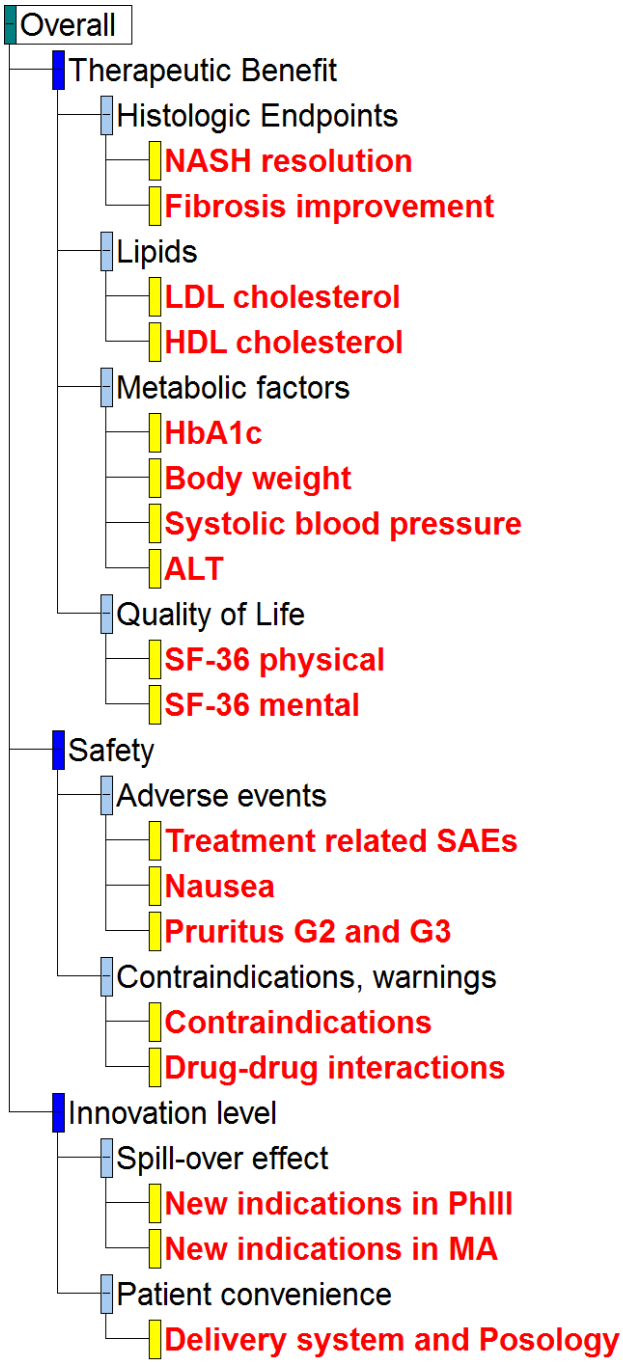


Round 2



c)

Round 1



Round 2