Supplemental Table 1. Biochemical parameters during clinical trials of combination therapy
(italics = desiccated thyroid extract)

Reprinted with permission from Taylor & Francis Ltd. From Jonklaas: Risks and safety of combination therapy for hypothyroidism. Expert Rev Clin Pharmacol. 2016;9(8):1057-1067.

Authors	Year	Serum TSH (mIU/L) in T4/T3 group	Serum TSH (mIU/L) in LT4 group	Diff ?	Serum FT4 (ng/dL) in T4/T3 group	Serum FT4 (ng/dL) in LT4 group	Diff ?	Serum T3 (ng/dL) or FT3 (pg/mL) in T4/T3 group	Serum T3 (ng/dL) or FT3 (pg/mL) in LT4 group	Diff ?
Appelhof 10:1	2005	0.35	0.64	N	1.02	1.18	Y	119 (T3)	111 (T3)	Y
Appelhoff 5:1	2005	0.07	0.64	Y	1.00	1.18	Y	143 (T3)	111 (T3)	Y
Bunevicius	1999	0.5	0.8	N	1.8	2.3	Y	117 (T3)	87 (T3)	Y
Bunevicius	2002	0.47	0.45	N	0.96	1.63	Y	246 (T3)	227 (T3)	N
Clyde	2003	2.0	2.1	N	0.8	1.2	Y	135 (T3)	87 (T3)	Y
Escobar-Morreale	2005	2.56	1.95	N	1.3	1.6	Y	3.2 (FT3)	3.3 (FT3)	N
Fadeyev	2010	1.9	2.4	N	n/a	n/a	n/a	n/a	n/a	n/a
Kaminski	2016	0.64	0.19	N	1.03	1.64	Y	98 (T3)	103 (T3)	N
Nygaard	2009	0.76	0.99	N	↓ FT4 index	FT4 index	Y	↑FT3 index	FT3 index	Y
Rodriguez	2005	5.6	2.7	N	↓ total T4	total T4	Y	99 (T3)	80 (T3)	Y
Saravanan	2005	2.28	0.73	Y	1.07	1.52	Y	2.5 (FT3)	2.4 (FT3)	N
Sawka	2003	1.8	1.7	N	0.81	1.38	Y	3.1 (FT3)	2.9 (FT3)	Y
Siegmund	2004	0.5	1.5	Y	1.56	1.62	N	2.9 (FT3)	3.2 (FT3)	N
Valizadeh	2009	2.5	2	N	↓ total T4	total T4	Y	164 (T3)	132 (T3)	Y
Walsh	2003	3.1	1.5	Y	11.4	15.6	Y	2.3 (FT3)	2.4 (FT3)	N
Hoang	*2013*	*1.67*	*1.3*	*Y*	*0.85*	*1.36*	*Y*	*138 (T3)*	*89 (T3)*	*Y*
↓ = decreased, ↑ = increased										

Supplemental Table 2. Outcomes of clinical trials of combination therapy
(italics = desiccated thyroid extract), grey shading indicates primary trial outcome(s)

Reprinted with adaptations with permission from Taylor & Francis Ltd. From Jonklaas: Risks and safety of combination therapy for hypothyroidism. Expert Rev Clin Pharmacol. 2016;9(8):1057-1067.

Authors, (Primary Outcome)	Year	Body weight /BMI	Lipid profile	BP	Bone turnover markers, BMD	Cardiac monitor-ing	Neuro-cognitive measures	QOL, mood, measures	Sig patient preference for LT4/LT3 over LT4*
Appelhof (preference)	2005	↓ combo 5:1	cholesterol ↓ combo 5:1, 10:1	No diff	Osteocalcin, alk phos ↑ combo 5:1	Pulse ↑ combo 5:1	No diff	No diff	Yes, 41% in 10:1, 52% in 5:1
Bunevicius (neurocog, QoL/mood [assumed])	1999	n/a	No diff	No diff	n/a	Pulse ↑ combo	↑ combo	↑ combo	Yes, 61%
Bunevicius (neurocog, QoL/mood [assumed])	2002	No diff	n/a	n/a	n/a	Pulse, echo parameters No diff	No diff	No diff (tendency ↑ combo)	n/a
Clyde (neurocog, QoL/mood)	2003	No diff	No diff	No diff	n/a	Pulse No diff	No diff	No diff	n/a
Escobar-Morreale (neurocog, QoL/mood, preference)	2005	No diff (BMI)	No diff	No diff	DPD No diff	Pulse ↓ combo	↑ combo (some parameters only)	No diff	Yes, 69%
Fadeyev (unclear)	2010	No diff	LDL chol ↓ combo	n/a	DPD ↑ combo, BMD no diff	Pulse No diff	n/a	n/a	n/a
Kaminski (QoL/mood [assumed])	2016	No diff	No diff	No diff	n/a	Pulse ↑ combo	n/a	No diff	n/a
Nygaard (QoL/ mood)	2009	No diff	n/a	n/a	n/a	n/a	n/a	↑ combo	Yes, 49%
Rodriguez (fatigue)	2005	No diff	n/a	No diff	n/a	Pulse No diff	No diff	No diff (no diff for fatigue)	n/a
Saravanan	2005	No	No diff	No	n/a	Pulse No	n/a	↑ combo	No (likert

(QoL/ mood)		diff		diff		diff		(3 mons only)	scale used)
Sawka (QoL/ mood)	2003	n/a	n/a	n/a	n/a	n/a	n/a	No diff	n/a
Siegmund (neurocog, QoL/mood)	2004	n/a	No diff	No diff	n/a	Pulse No diff	No diff	No diff	n/a
Valizadeh (neurocog, QoL/mood [assumed])	2009	No diff	No diff	No diff	n/a	Pulse No diff	n/a	No diff	n/a
Walsh (neurocog, QoL/mood [assumed])	2003	No diff	↑ combo	No diff	Alk phos, DPD, No diff	↓ combo	No diff	No diff	No
Hoang (neurocog, QoL/mood)	*2013*	↓ *combo*	*No diff*	*No diff*	*n/a*	*Pulse No diff*	*No diff*	*No diff*	*Yes, 49%*

↓ = *decreased,* ↑ = *increased, DPD* = deoxypyridinoline, sig = statistically significant, n/a = not assessed or informally assessed

**A meta-analysis by Akirov et al. found that preference for LT4/LT3 did not differ from preference predicted by chance*

Supplemental Table 3. Canonical and non-canonical effect of thyroid hormone

	Type 1 – nuclear canonical	Type 2 – nuclear without DNA binding	Type 3 – cytosolic TR	Type 4 – cell membrane not requiring TRs
HPT axis feedback	+ (TRβ)			
Hearing	+ (TRβ)			
Browning of white adipose tissue	+ (TRβ)			
Bone development	+ (TRα)			
Direct regulation of target gene expression	+ (TRα/TRβ)			
Vasodilatation			+ (TRα)	
Hepatic and serum triglyceride concentration			+ (TRβ)	
Tumor growth				+ (αvβ3)

Supplemental Table 4. Symptoms of importance to patients versus physicians

Adapted with permission from table 3 of Watt et al. Thyroid 17 (7): 647-654, 2007. http://doi.org/10.1089/thy.2007.0069

Important according to:		
Patients only	Both groups	Clinicians only
Being slow* Bags under the eyes* Getting upset* Palpitations Globulus sensation Dyspnea Clearing throat often*	General fatigue Cold intolerance Physical fatigue Hypersomnia Weight increase Weight dissatisfaction Mental fatigue Constipation	Impaired memory Limit daily activities Hoarseness Difficulty concentrating Depression Dry skin Attention problems

Supplemental Table 5. Comparison of number of patients in prior studies compared with the number required for adequate power for a QoL outcome, depending on effect size

Studies	**Number of patients in study**	**Number required for an effect size of 0.5**	**Number required for an effect size of 0.3**	**Number required for an effect size of 0.5 if only 20% of patients were expected to benefit from therapy**
Cross-over studies				
Bunevicius 1999	33	40	107	944
Walsh 2003*	110			
Siegmund 2004	26			
Rodriguez 2005	30			
Escobar-Morreale 2005	28			
Nygaard 2009	59			
Parallel group studies				
Clyde 2003	44	128	352	3142
Sawka 2003	40			
Saravanan 2005*	697			
Appelhof 2005	141			
**adequately powered studies if all patients expected to benefit from therapy*				

Supplemental Table 6. Secondary efficacy and safety outcomes for a combination therapy trial

Category	**Secondary efficacy outcomes**	**Safety measures**	**Consider for sub-group, proof of concept**
Metabolic	Body weight		
	Lipid panel		
			Body composition
			Resting energy expenditure
Cardiovascular	Resting heart rate		
		Cardiac arrhythmia monitoring	2-week cardiac monitoring
Cognitive	Fluid cognition tests		
Musculoskeletal	Bone biomarkers (C-telopeptide)		
			DXA
Cancer risk or progression		Breast cancer screening	
Side effects			Peak T3 levels after LT3 dosing
		Hyperthyroid symptoms	
		Adverse events	

Supplemental Table 7. Consensus Statements Organized by Topics (topics 1-10), with selected comments, as follows:

a) selected comments made by authors during the voting process retained in red font to provide additional information regarding the discussion and debate. *Comments are from all the rounds of voting and may have contributed to changes being made in the consensus statement. Votes may also have been changed after the comments were made.*

b) Selected comments made by ATA and ETA members during the open member review period in blue font.

Topic 1 **Local control of thyroid hormone action, type 2 deiodinase polymorphisms…**	Comments or suggestions
1.1 Future trials of combination therapy in humans should consider including genotyping for the Thr92AlaD2 polymorphism, and should be adequately powered to study the effect of this polymorphism on study outcomes. Degree of Consensus 100%	This is expensive and potentially difficult to do. I disagree with powering based on genotype testing but agree with powering based on known population prevalence. I agree on principle, but I disagree due to impracticality. I am concerned that this recommendation will prevent trials from getting funded unless they have sufficient power to test both the effect of combo vs mono and the effect of polymorphism vs. no polymorphism. If someone can show me a power calculation that this will not require an enormous study, then I will vote yes for should consider but no for should. I would hate to see someone on a review panel fault a study that acknowledges that it is underpowered to answer that question. If looking at PROs, This will require thousands if powered to the Thr92Ala polymorphism. A better statement may be: clinical trials are needed examining the effect of the Thr92Ala polymorphism on patient outcomes. Agree with statement ideally. Feasibility considerations are in part dictated by the intensity of resource use for the genotype testing, so maybe if the test is done accurately, easily and cheaply on a large scale, then may be feasible. Other genetic backgrounds involved in the resistance to exogenous thyroxine (RETH) other than Thr92AlaD2 polymorphisms might be acknowledged, especially in patients with a particularly severe form of the phenotypic spectrum, as defined by iatrogenic thyrotoxicosis followed by thyrotropinemia,

	after failure of thyroid function by e.g. thyroidectomy (DOI:10.1089/thy.2019.0825).

Topic 2 **Non-classical actions of thyroid hormone**	Comments or suggestions
2.1 Consideration should be given to assessment for effects of thyroid hormones that may be manifest via non-canonical as well as canonical pathways (e.g. triglyceride levels and cardiac function) in future trials of combination therapy. Degree of Consensus 83%	The biological effects of thyroid hormone with clinical relevance are well known. The discovery of non-canonical pathways (which has not been reproduced by other laboratories) explains some of the mechanisms involved but does not affect clinical decision making; for example, T3 accelerates heart rate, regardless of the pathway. I like the comment above, because it is very reasonable assuming that it does not matter whether T3 acts via canonical or noncanonical paths. I agree and would not expect differences. However, the parameters collected will be collected anyway (lipids incl. TG; heart rate, possibly echocardiography, blood pressure), so those effects will be studied anyway, regardless of the pathway. It may be more important to consider the underlying mechanism when analyzing the results. This is interesting, but appear too premature to form the basis for a recommendation for near-future trials. A better statement may be: evaluation for the effects of thyroid hormones that may manifest via non-canonical and canonical pathways is needed in future clinical trials and in basic science/translational research. If less than physiological replacement is used e.g. non-slow release T3, this may have different effects on canonical vs non-canonical signaling . This would be best assessed in small proof-of-concept targeted studies rather than in a large one.
2.2 Consideration should be given to assessment for effects of thyroid hormones that may operate by non-TR mediated pathways (e.g. cancer progression) in future trials of combination therapy. Degree of Consensus 25%	*(see also comments in relation to 2.1 above)* Interesting that so many disagree here. 'Future trials of combination therapy' could include therapy of patients developing hypothyroidism on cancer therapy… In those patients, samples sizes would me much smaller… I assume that everyone thought of trials in hypothyroid patients that are otherwise healthy, but trials could also specifically study hypothyroid cancer patients for progress of their cancer… It is fine to consider this, but realistically the kind of sample sizes and

	duration of follow-up needed to properly examine cancer risk may not necessarily be feasible for clinical trials. Could be complemented by population-based/administrative database research, prospective cohort studies, registries. Suggested alternate wording: More research is needed examining the effects of thyroid hormones that may operate by non-TR mediated pathways (e.g. cancer progression). We do not have power to look at cancer progression in RCTs. A six- or twelve-month trial would be inadequate to provide the information. Might be very interesting to study hypothyroid and also LT-4 overtreated patients regarding cancer progression.

Topic 3 **Thyroid Hormone Transporters and CNS Levels of Thyroid Hormone**	Comments or suggestions
3.1 A consideration for future trials of combination therapy in humans is that they could be adequately powered to study the effect of polymorphisms in thyroid hormone transporters (e.g. MCT8, MCT10, OATP1C1) on study outcomes. Degree of Consensus 100%	I assume that this statement is based on the Carle Eur Thyroid J 2017 paper? To support or justify such a statement, the available evidence for effects of polymorphisms in humans should be explained first. We don't even know the transporter equipment of many cells and its consequences. Therefore we need to be open for additional transporters and their polymorphisms. The more knowledge the better… If transporter polymorphisms alter physiology, we should know and substitution studies may be one way to clarify their physiological relevance. Therefore I changed to the 'agree' group. This is expensive and potentially difficult to do. I disagree with saying should, but saying could is acceptable. Studies should be powered based on known population prevalence. Clinical trials are needed examining the effect of the polymorphisms in the thyroid hormone transporter on patient outcomes would be my suggested alternative wording. Not clear how large the sample will need to be for assessment of the effect of multiple polymorphisms – may be too large for a single realistic trial. Once multiple polymorphisms are taken in account it would be

	extremely difficult to sort out the effect size of a single one. Most likely secondary analyses may generate "risk scores". I agree on principle, but I disagree due to impracticality. I am concerned that this recommendation will prevent trials from getting funded unless they have sufficient power to test both the effect of combo vs mono and the effect of polymorphism vs. no polymorphism. If someone can show me a power calculation that this will not require an enormous study, then I will vote yes for "should consider" but no for "should". I would hate to see someone on a review panel fault a study that acknowledges that it is underpowered to answer that question.

Topic 4 **Selection of participants for combination therapy trials**	Comments or suggestions
4.1 Patients who do not report relief of their symptoms with LT4 therapy should specifically be recruited for combination therapy trials. Degree of Consensus 75%	Wouldn't this introduce a strong selection bias? Would it not be reasonable to thoroughly characterize these patients first before starting an intervention? All considerations regarding DIOs, transporters, modes of action assume that there could be a physiological basis for dissatisfaction. I would therefore first see whether there is a difference between satisfied and unhappy patients with regard to transporter and DIO polymorphisms, thyroid function tests etc. Sort of 'characterization before intervention'. Such a study would be MUCH easier, no dose adjustment, no treatment, much easier to obtain ethical permission etc. If there indeed is a difference, then the combination may be much better justified. Suggested alternative wording would be clinical trials are needed examining whether combination therapy improves symptoms in LT4-treated patients with persistent hypothyroid symptoms. Symptoms could be evaluated at baseline using PROs (combine with 4.2). This assumes that psychological parameters are the only important outcome measures. However, it may reduce the sample size required to assess this outcome. The "two-step approach" [first characterize (DIOS, transporters, metabolomics) the patients – is there a difference between satisfied and unhappy patients? and then a RCT in the group with persistent complaints] is now being planned in the Netherlands. But I am not sure if dissatisfied patients should specifically be recruited in ***every*** trial…

	A robust biological biomarker the RETH phenotype (doi: 10.1089/thy.2019.0825), seems to be a decreased T3/rT3 ratio. This biomarker opens the possibility of performing two-arm intervention trials distributing patients on *objective* biological data under LT4 treatment (say, decreased or normal ratio), and not on (subjective) patients complaints or QoL questionnaires. Practicality/cost of these determinations can be discussed, but the current feasibility of that approach might be acknowledged. Patients who do not report relief with levothyroxine treatment and have a total T3 in the lower 50th percentile should be recruited for combination therapy trials.
4.2 One or all of several previously-validated thyroid-related quality of life questionnaires should be used to assess the baseline dissatisfaction to be used as an inclusion criterion. Degree of Consensus 100%	Suggested alternative wording would be clinical trials are needed examining whether combination therapy improves symptoms in LT4-treated patients with persistent hypothyroid symptoms. Symptoms could be evaluated at baseline using PROs (combine with 4.1). Important that studies are powered on a single, pre-selected primary outcome.
4.3 Patients should be treated with at least 1.2 mcg/kg/day of LT4 in order to be eligible. Degree of Consensus 100%	One could argue that subclinical hypothyroidism was a group of interest – patient with a slight increase in TSH – given T4 could induce lower T3 and then given a more pronounced fall in QoL. This is a proxy for lack of endogenous secretion of thyroid hormone, not perfect, but "good enough" for large trials. I am not sure that this relationship is constant in elderly patients with low muscle mass. Could stratify patients according to dosage relative to weight but not sure that I would make it an inclusion criterion. I realize you are trying to get at residual thyroid function, but maybe could analyze that by addressing with stratification or maybe a secondary analysis. This represents 100 mcg/day of T4 in n 83 Kg individuals. This seems reasonable and will include the majority of patients with subclinical hypothyroidism where optimization of treatment with T4 has been attempted. Higher doses of T4 are required to show if there is an "inhibitory effect" on D2 activation by increased FT4/FT3 level. Athyreotic patients under LT4 monotherapy should be included in trials.

	My question is there a difference between postoperative hypothyroidism and hypothyroidism due to thyroid atrophy or chronic thyroiditis in regard to LT4/LT3 therapy?
4.4 Patients who have low baseline serum total T3 levels while taking LT4 monotherapy should be included in trials, and results could be stratified according to the change in trough total T3 levels achieved with combination therapy. Degree of Consensus 50%	T3 immunoassays exhibit high variability and might super-estimate results at lower T3 levels; in addition, T3 levels are greatly affected by caloric/carbohydrate intake; patients on caloric restriction could have low serum T3 and qualify to a trial, introducing more variables; serum T3 seems to be acceptable when looking at large populations, which dilutes these interfering factors; on an individual basis I worry about using serum T3. Serious question: Is there a trough T3 on T4 monotherapy? We state that T3 levels achieved on T4 are, in fact, stable. Simplify wording: Inclusion of patients with low trough T3 levels on LT4 monotherapy should be considered in future clinical trials of combination therapy. T4/T3 ratios are likely to be more important than T3 alone. Rather than using low T3 as an inclusion criterion, this parameter could be initially used as an explanatory one in the analyses. After all patients do not complain of being affected by low T3 levels. I think serum T3 is an interesting parameter in post-hoc analyses, but I don't agree that patients with low serum T3 should specifically be targeted for trials. It is disappointing that, and no doubt many hundreds of patients whose symptoms did not completely resolve will be disappointed, Topic 4.4 will not be included in the list as it only achieved a 50% consensus. This is a trial that many, many patients have been waiting to see. If this trial was undertaken I am sure that the T3 immunoassay variability problem could be overcome and patients on caloric restriction could be excluded. Slow release T3 is available from at least one compounding pharmacy which would then presumably, stop the problem of T3 troughs. It's very difficult to stratify by FT3 given the age-related decline in FT3 reference ranges and diurnal variation in FT3. Plus, lower FT3 values reflect degrees of comorbidity as well as thyroid status. This would be unhelpful in clinical practice as you won't be able to predict the change in FT3 up front.

Topic 5 **T3/T4 Dose Equivalence – Clinical and Trial Data**	Comments or suggestions
5.1 Future combination therapy trials should incorporate measurement of trough levels of both serum FT4 and total T3 (for example, as a nested pharmacokinetic study in a representative small sub-group). Degree of Consensus 92%	Some comments on the limitations on the T3, FT3 methods would be relevant. I do not consider this a requirement for clinical conclusions or generalizability of a future trial. "Nested" PK studies in a small but representative group of the study population could be considered. The rationale for nesting depends on if you would use these measures to actually titrate doses in the trial or not. If you are, then everyone would need to have the measures to guide the therapy. However, if you are not using the levels to guide dose adjustment, nested pharmacokinetic study is fine.
5.2 Future combination therapy trials should incorporate measurement of peak levels of serum total T3 (approximately 1.8-2.5 hours after LT3 administration) as a nested pharmacokinetic study in a representative small sub-group. Degree of Consensus 83%	Some comments on the limitations on the T3, FT3 methods would be relevant.. I do not consider this a requirement for clinical conclusions or generalizability of a future trial. "Nested" PK studies in a small but representative group of the study population could be considered. This may not be feasible for every measurement in every patient in a very large trial, could do in a subset of patients or subset of measures. However, it really depends on how you are titrating the doses….may be essential if you are using those specific measures for dose titration within a trial. The mechanism of T3 and T4 combined therapy in regulating TSH level needs to be explored. T4 metabolism is slow, half-life is a few days. The half-life of T3 is shorter, even if given 2-3 times a day, T3 concentration fluctuates significantly within 24 hours. Therefore, it is necessary to detect the peak level of T3 to evaluate the follow-up efficacy.

Topic 6 **Target T3 and TSH levels and Slow Release T3**	Comments or suggestions
6.1 The goal of future LT4/LT3 combination studies should be to achieve a	Not a clinical necessity. We state in this manuscript that it is uncertain whether peak or trough

physiological FT3/FT4 ratio. Degree of Consensus 67%	T4/T3 ratio should be used. How can we advise to achieve a physiological ratio when we don't know when that should be measured? This again indicates that new trial make sense only with slow release T3 (or T3S) and stable serum concentration. Is this really the main goal? I am not sure. Could argue symptom relief is more important to patients, regardless of physiologic ratio? Improving psychological symptoms without a physiological T4/T3 ratio may compromise effects in other tissues. There is some concern about supraphysiological levels being associated with negative effect on safety i.e. compromise effects in other tissues….do we have published evidence that the ratio is better predictor of adverse consequences than either TSH or free thyroid hormone levels on their own in general or in context of combination therapy? Since there is variation within normal population, I think this goal is neither feasible nor clinically relevant. Perhaps rephrase: One of the goals of future…(a physiological FT3/FT4 ratio is certainly not the main goal as long as we don't have adequate slow-release preparations). It is disappointing that Topic 6.1 has not been considered for inclusion. Slow release T3 is available from at least one compounding pharmacy which would then presumably, stop the problem of T3 troughs. I think it is not possible identify a unique FT3/FT4 ratio. The main goal should be the relief of symptoms.
6.2 If non slow-release LT3 therapy is used, it should be given at least twice daily. Degree of Consensus 100%	Given the very good points above, should we not better discourage trials without slow-release T3? But I support 'at least twice daily' with the explanations given in the text. Unless there is clear evidence of lack of clinically relevant nongenomic effects of T3. Yes ideally, but some patients will not do this. Not sure if I would insist on it. Depends on if you think the outcome effects are dependent on not having fluctuation levels. We assume so, but not really sure. I think that T3 twice daily administration is a good method for future

	clinical trials. I disagree to discourage future clinical trials because Slow Release T3 is not available. I believe the physiology dose of T3 is actually three times daily, which should be written as the goal, even patient can only do twice a day.
6.3 The use of slow release T3 preparations is desirable in future trials of combination LT4/ LT3 to achieve physiological levels of thyroid function. However, no approved slow release T3 therapies are available at this time. Degree of Consensus 100%	

Topic 7 **Psychological and Quality of Life Measures**	Comments or suggestions
7.1 If a PRO is used as a primary outcome in clinical trials, the measure should have well-documented content and validity for thyroid related QoL as well as responsiveness to change. Degree of Consensus 100%	
7.2 Future studies need to be appropriately powered for PROs as primary outcomes based on the primary endpoint on an effect size of at least 0.5, and preferably 0.3. Degree of Consensus 100%	If the primary endpoint is satisfaction/well- being. In other words, a proof-of-concept study powered for physiology outcomes could use a PRO as a secondary/exploratory endpoint. Should we be more specific about which PRO?
7.3 ThyPRO-39 is favored as a primary QoL endpoint for the study. Degree of Consensus 100%	But the secondary outcomes to document effects on other tissues especially with non-physiological replacement are important.
7.4 Patient preference should	Reserve as a secondary outcome.

be included as a secondary trial outcome. Degree of Consensus 100%	Secondary outcome ok with me as may not be validated scales at present but patients are interested in it. A fundamental unanswered question is whether the "stimulatory" effect T4/T3 combination Rx, DTE, and specifically T3 is what leads to patient preference. Studies on the impact of T3 as adjunctive psycho-pharmacotherapy are mixed. Nonetheless this might be addressed by taking those without thyroid dysfunction and assess the impact of combination "Rx" on them. Case selection could include those who feel well and those who don't; those with + antithyroid antibodies and those who don't...
7.5 A qualitative study should be considered to explain patient preferences for thyroid hormone formulations. Degree of Consensus 75%	Much of this work (Prior statement: "A qualitative study should be considered to determine which items are important to patients. If these items are not contained within existing PROs, they could be added") has already been done in development of the ThyPRO. Could look at why patients prefer combination therapy over LT4 or vice versa, which is not necessarily covered in ThyPRO development. Good idea. Preferably using strategic sampling, selecting participants immediately after they have provided their preference rating. There is quite some literature on the symptoms and complains associated with hypothyroidism, their weight at the individual's level varies immensely. Some patients organizations have organized qualitative studies, largely unpublished. A formal qualitative study will take a lot of time and delay future quantitative studies. This is not my area of expertise… However, such a study appears to do no harm as the result will either be that the items are either already included in existing PROs or not. Importance to patients may be a value in itself as future trials will show whether these items were improved or not with combination therapy, thus improving advice to patients. Maybe especially so in the absence of changes in those items.

Topic 8 **Biological Outcomes, Biomarkers and Safety Measures**	Comments or suggestions

8.1 Metabolic efficacy outcomes in future trials should include body weight and lipid panel. Resting energy expenditure should be considered for study in a nested sub-group. Degree of Consensus 92%	I am not completely convinced to include REE, as we don't know the individual ideal REE. In addition when using indirect calorimetry there is an intra-individual variation of ~3%... REE measurement is extremely difficult, I would suggest to perform this measurement in a "nested" subgroup or in a proof-of-concept study. I do not think that REE should be a must – the expected changes are small, and I will be costly and time consuming to included it a large trial. Not necessarily resting energy expenditure. REE is not a standard clinical measure and is not an essential measure. Weight is an appropriate metabolic efficacy surrogate. May want to reword beginning of sentence as: If metabolic efficacy outcomes are studied in future clinical trials, these ?could ?should include: weight, lipid panel, and resting energy expenditure. REE is related to free fat mass (FFM) and to body cell mass (BCM), correlate FT4/FT3 with REE is important especially in overweight patients where FFM and BCM might be very different respect to normal weight patients. Possibly include hours of sleep into metabolic efficacy measures.
8.2 Cardiac efficacy outcomes in future trials should include resting heart rate. Degree of Consensus 100%	If cardiac efficacy outcomes are studied in future clinical trials, these ?could ?should include the following...
8.3 Cognition efficacy outcomes should include fluid cognition testing. The NIH Toolbox cognitive battery is a viable option. Degree of Consensus 92%	I am not familiar with the use of NIH Toolbox cognitive battery (the cost of using it?– is it possible to use it as an endocrinologist or nurse or do you need a psychologist or a specialist in neurology to use it?) – suggest rephrasing this to a suggestion of including fluid cognition testing i.e. using the NIH Toolbox cognitive battery. If cognitive efficacy outcomes are studied in future clinical trials, these ?could ?should include the following. Maybe as a secondary outcome, but I am concerned about the excessive questionnaire burden for patients. Translation to other languages is an issue with this instrument (very expensive).

8.4 Musculoskeletal efficacy outcomes in future trials should include a bone biomarker (e.g. C-telopeptide), and should consider measurement of bone density using DXA scan if the trial is 12 months or longer in duration. Degree of Consensus 92%	Do you mean DXA to measure body composition or for osteoporosis? It is important to measure risk of osteoporosis when patients are treated for a long period of time. – not in a shorter trial of < 12 month . Perhaps HR-pQ CT or other more sensitive estimates. This and possibly some other parameters are only reasonable when the trial duration is long enough for changes to manifest. For example, how long does it take for T3 to induce measurable bone loss? If musculoskeletal efficacy outcomes are studied in future clinical trials, these ?could ?should include the following. DXA seems only relevant for very long-term studies (e.g. 1-3 years).
8.5 Safety monitoring should incorporate measurement of thyrotoxic symptoms, hypothyroid symptoms, and adverse events. Degree of Consensus 92%	Consider including hypothyroid symptoms as well. Safety monitoring ?could ?should include the following. This seems a bit vague – how can we capture this in a robust way? This would be included in the Adverse Events section of the trial.
8.6 Safety monitoring should incorporate cardiac monitoring with ECG at baseline and 3 month intervals. Cardiac rhythm monitoring of longer duration could be considered in a nested sub-group. Degree of Consensus 83%	2-week cardiac monitoring could be a limitation for making the study – suggest rephrasing to a suggestion of cardiac monitoring. Not sure that such extensive monitoring is worthwhile (necessary/can pick up complications anyway). When? At the beginning or at the end of the trial? For example, people that develop atrial fibrillation due to hyperthyroidism and hot nodule, for how long has this slowly developing hyperthyroidism had time to ultimately result in AF? Would it be more reasonable to monitor at the beginning or rather later in the trial? This seems overreaching, , I would suggest to perform this measurement in a “nested” subgroup or in a proof-of-concept study. Not sure of justification. Any occurrence of A-F is an important safety outcome and should be detected. No. If longer monitoring i.e. 24 hours or longer, this should be in all participants as the number with A-F or arrhythmias especially in younger people will be small. Seems to be significant disagreement. The way this stands, it appears

	that you are routinely recommending cardiac rhythm monitoring in all trial patients, without actually saying type, duration or timing of monitoring. From a practical standpoint, there is a big difference between resting ECG at baseline and end vs. wearable monitoring device. Nested could be one way to do it, and may want to stratify by age and other risk factors for AF, particularly if prolonged monitoring,. Resting ECG may be less of an issue to quibble about feasibility, but then get into whether it is sensitive enough to pick up transient changes.. If do nested comparison, need to pay attention to stratifying by age and other risks for A fib. But I think the suggestion of rewording is a good idea to soften the statement and get more authors in agreement. "Safety monitoring could incorporate cardiac rhythm monitoring". If financially possible, might be more useful to have continuous cardiac rhythm monitoring the first 1-2 weeks on combination therapy treatment.
8.7 Pilot trials are needed to explore additional outcomes of secondary importance as well as relationships between variables. Such studies may be conducted within a larger trial. Degree of Consensus 50%	For example we cannot expect to have all the patients undergoing energy expenditure measurements in the whole room calorimeter, but a small group could produce important data. These can be done in parallel to guide future studies, otherwise the primary studies will be significantly delayed. Will take a lot of time, while some/much of this information is already available in the literature… Here, the conflict between the voters seems to be between 'preliminary' and 'subgroup in parallel'. Similarly, the main text starts with the suggestion of preliminary studies (first sentence) and then suggests 'subgroups' (second sentence). The main text is undecided between the two options, whereas '8.7' indicates preliminary studies. The solution could be to allow/suggest both and explain both in the main text. I do not agree that these should be run in parallel. They should precede a large trial. No. Pilot studies should be done prior to the main trial and the participants could be incorporated. If run in parallel this is too late, and you might as well use the main trial to inform the next study. I would avoid this. Does seem to be a suboptimal compromise and

	now an unnecessary statement. Energy expenditure could be studied by the corrected Mifflin formula, allowing a measurement of larger sample in the contest of a large trial (Am J Clin Nutr 1990 Feb;51(2):241-7. doi:10.1093/ajcn/51.2.241).

Topic 9 **Trial Design Considerations**	Comments or suggestions
9.1. A future combination therapy trial should be randomized, placebo-controlled and double blinded. Degree of Consensus 100%	Subjective outcomes are highly susceptible to the quality of blinding. You need to discuss that true blinding to LT3 therapy may be impossible due to short-term effects on heart rate, facial flush etc: "the hit" mentioned by your patient participant. This is specifically mentioned in the EU LT3 pack insert (tachycardia, loose stool, sleep disturbance). It's obvious that dissatisfied people who feel these physiological effects will feel the treatment is doing them more good. Also may have been unblinded by smell or taste of pork.
9.2. A future combination therapy trial should be at least a year in duration, with interim outcome assessments at 3 and 6 months. Degree of Consensus 80%	To detect "slow onset" biological effects. Here we are considering the equivalent of a phase 3 trial, i.e. we are not discussing proof-of-concept trials. Duration depends on the primary and secondary outcome measures, unnecessary to limit this to at least 1 year. Of course, the longer the better; however, 1 year will increase the price tag of any study by a lot; in the studies including placebo the effects seem to wane after 3-4 months; thus, minimum of 6 months could be informative. Agree, but may limit inclusion rate, due to patients not willing to wait a whole year before receiving combination therapy, if they think it is beneficial.
9.3 A future combination therapy trial should incorporate a parallel design. Degree of Consensus 100%	Agree in general for a definitive trial, especially for QoL and safety outcomes– but crossover could be ok for a proof of concept surrogate outcome or preference primary outcome study (for preference study it is nice to be able to do blinded evaluations sequentially within individuals, so individuals can compare the therapy. Here we are considering the equivalent of a phase 3 trial, i.e. we are not discussing proof-of-concept trials. Challenges an intention to include preference as an outcome, though.
9.4 A future combination therapy trial should consider	We should aim first for physiological replacement.

incorporating an arm being treated with DTE, in addition to the LT4 and LT4/LT3 arms. Degree of Consensus 67%	Patients really want research on DTE. Not sure if you want to say this should be a head to head comparison that is adequately powered to detect differences. A three-arm trial is desirable, but that would entail extensive preliminary work to demonstrate near-equivalence of the formulations. Will generate problems (e.g. reproducibility) because the intervention is not clearly defined in terms of composition, content…. Would prefer "could", rather than "should", though. The topic of T4 vs. T4/T3 is already so complex and unresolved. I would address DTE once the first question is resolved. By the way, DTE plays almost no role in parts of Europe. Although I do not favor the use of desiccated thyroid extract (DT), more extensive consideration should be given to a trial involving DT because of the resurgence of popular interest in its use and the paucity of data comparing it to LT4. Treatment with desiccated thyroid extract should, we feel, have been included especially since many patients are taking this medication due to the fact that liothyronine is so expensive and will probably remain so. We know that many patients feel so much better on this treatment than on levothyroxine and it should at least be put forward for future trials. I would address DTE once the first question (LT4 vs LT4/LT3) is resolved.
9.5. It is important for future trials to be pragmatic and include patients with managed, stable comorbidities, so that the results are generalizable to the hypothyroid patient population. Degree of Consensus 90%	I can see a large and diverse population which includes patients with comorbidities only in an effectiveness trial. The latter can be conducted only after a series of proof of concept preliminary trials aimed to determine the dosing, frequency, and point estimate. The risk of starting with a "all-comers" is to dilute the internal validity of the study and to lose statistical power. Keep for a later study, as this may complicate assessment of biological outcomes (e.g. CV parameters). Yes but may need to say something about doing this with attention to safety.

Topic 10	Comments or suggestions

Incorporation of patient experiences	
10.1. A 2 x 2 factorial design randomized controlled trial, randomizing patients to either combination therapy or LT4 and to either a lifestyle intervention (e.g. education, diet, exercise, or a combination) or no lifestyle intervention, would inform the understanding of pharmacologic and non-pharmacologic intervention effects on patient experiences of their therapy. Degree of Consensus 50%	There is no basis for thinking that lifestyle intervention improves symptoms of hypothyroidism. Depends on design: Yes, if mono- and combination therapy are given in a randomized way to patients with lifestyle modification. There needs to be placebo control for combination therapy. An unblinded lifestyle arm would be an inadequate control group. There could be consideration of lifestyle vs standard of care in LT4 users at TSH goals. I am also not sure that it is necessarily appropriate to lump all lifestyle interventions together as it depends, in part, what symptom is targeted. For example, for brain fog – cognitive behavioral therapy, exercise but not necessarily diet may be helpful. Weight loss – diet, exercise. Fatigue – exercise. Also what do you mean by "extensive" lifestyle modification and whether education about hypothyroidism should be standard care for all patients in all arms. What kind of education do you mean? Agree with concept but could be more clear. I think you are trying to suggest a factorial design, which is appropriate. Also, for lifestyle arm, you have a combination lifestyle intervention, but may actually want to do something like individual interventions, so would not want to say that MUST do diet/exercise/education. That is a good combo but actually does not allow you to figure out what is helpful and could be hard to implement all 3 at same time, particularly if trying to do in standardized way, properly. You have a lot of conflict here. Not sure this can be resolved. May need to remove or change or just acknowledge lack of consensus and explanation of reasons. The problem is that this was actually asked for by the patients. Could state something more general like the following: The effect of lifestyle modification (e.g. diet, exercise), behavioral therapy, and supplements needs to be studied, with respect to improving health outcomes of treated hypothyroid patients. This vague statement could encompass potential factorial designs, etc. I am not sure how patients will feel if you completely delete this topic as they raised it repeatedly in session and this needs to be patient-

	centered. Lifestyle modification against a background of LT4 only? Then there should be an additional arm for lifestyle modification against a background of combination therapy… So I don't think we should incorporate this in a consensus statement on LT4/LT3 combination therapy. I would reserve this for follow-up studies once the original question has been resolved. Yes, but this kind of study is very difficult to "blind" making it difficult to discern an attention/placebo effect from a specific effect of the intervention strategy. However, irrespective of the mechanism of action, it would be good to know that there is something that helps. In my view the priority is to show whether there is a biological need for T3 replacement, and how to deal with "dissatisfaction" per se by other means is a separate question. It is a very complex study design and moreover it not easy to measure the compliance of each patients to the life style intervention.
10.2 The level of interaction between patient and physician should be considered as a factor affecting satisfaction with therapy in future trials, and should therefore be carefully standardized. Degree of Consensus 92%	Presumably, this will be balanced by randomization . Sounds good in principle but how do you measure that? Maybe some qualitative interviews? Shouldn't any communication in a trial be fairly well standardized? May want to train trial staff and have standardized approach for communication. I am not sure if you mean in a real life effectiveness study or an efficacy trial? Blinding may help here.
10.3 Fatigue/tiredness measures can be assessed in future trials using the composite scale of ThyPRO 39 (see topic #7 also). Degree of Consensus 92%	I think the ThyPRO-39 might be too crude a fatigue measure for such a trial. I would go with the Composite scale, if using ThyPRO-39 And use the full Tiredness scale from the 85-item ThyPRO (which is comprehensive and sensitive, also compared to standard measures). Include prior comment on composite scale. There are other dedicated general fatigue measures – can those be used? I wonder if you could combine this with other statement on ThyPRO in topic 7. OK but I do not think it is only way to measure, but can live with it as it would be efficient to evaluate this subscale if already using the ThyPRO for the primary outcome of QoL. So why don't the mental and physical fatigue scores of SF36 change

	in the T3/T4 combination trials. Are you implying that thyroid patients have a different sort of fatigue to patients with cancer, diabetes, heart failure, arthritis, for whom SF36 clearly works? The wording of the questions presumably cannot be so different.
10.4 Neurocognitive testing instruments selected for future trials should be tested to determine if they are responsive to changes in "brain fog" (see topic #8 also) Degree of Consensus 92%	This could perhaps better be rephrased as Summary statement: at this point, it is uncertain if… etc. Good idea, but not necessarily. OK but there are specific tests listed in 8.3 – do those meet the criteria needed for use? Also, maybe could just merge this with 8.3 as seems overlap concepts. Extensive neurocognitive testing is time consuming. I think it would be best included in a "nested" study/center with expertise.