CENTER FOR DRUG EVALUATION AND RESEARCH

APPLICATION NUMBER:

205598Orig1s000

STATISTICAL REVIEW(S)

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research
Office of Translational Sciences
Office of Biostatistics

STATISTICAL REVIEW AND EVALUATION

CLINICAL STUDIES- ADDENDUM

NDA/BLA #:	NDA 205598
Drug Name:	Macrilen (macimorelin)
Indication(s):	Growth-Hormone Stimulation Test (GHST) for diagnosis of adult growth hormone deficiency (AGHD)
Applicant:	Aeterna Zentaris
Date(s):	Stamp: 6/30/2017 PDUFA due date: 12/30/2017
Review Priority:	Standard
Biometrics Division:	II
Statistical Reviewer:	Anna Kettermann, Dipl. Math, MA
Concurring Reviewers:	Gregory Levin, PhD, Acting Team Leader
Medical Division:	Metabolism and Endocrinology Products
Clinical Team:	William Lubas, MD, Medical Reviewer Marina Zemskova, MD, Clinical Team Leader
Project Manager:	Meghna Jairath
Keywords:	diagnostic accuracy, diagnostic agreement, sensitivity (specificity)

Table of Contents

1 EXECUTIVE SUMMARY 5

2 INTRODUCTION 8

2.1 OVERVIEW 8
2.1.1 Indication 8
2.1.2 History of Drug Development 8
2.1.3 Specific Studies Reviewed 9
2.2 DATA SOURCES 9

3 STATISTICAL EVALUATION 9

3.1 DATA AND ANALYSIS QUALITY 9
3.2 EVALUATION OF EFFICACY 9
3.2.1 Study Design and Endpoints 10
3.2.2 Statistical Methodologies 13
3.2.3 Patient Disposition, Demographic and Baseline Characteristics 16
3.2.4 Results and Conclusions 20
3.3 EVALUATION OF SAFETY 31

4 FINDINGS IN SPECIAL/SUBGROUP POPULATIONS 31

4.1 GENDER, RACE, AGE, AND GEOGRAPHIC REGION 31
4.2 OTHER SPECIAL/SUBGROUP POPULATIONS 34

5 SUMMARY AND CONCLUSIONS 37

5.1 STATISTICAL ISSUES 37
5.2 COLLECTIVE EVIDENCE 39
5.3 CONCLUSIONS AND RECOMMENDATIONS 40
5.4 LABELING RECOMMENDATIONS 40

REFERENCES: 41

APPENDIX 42

Reference ID: 4188461

LIST OF TABLES

Table 1. List of all studies included in analysis 9
Table 2. Definition of agreement 14
Table 3. Baseline demographic characteristics: Age distribution (years) 16
Table 4. Demographic table: Race distribution 17
Table 5. Demographic Table: BMI distribution 18
Table 6. Demographic Table: Sex distribution 18
Table 7. Demographic Table: Region 18
Table 8. Demographic Table: Country 19
Table 9. Levels of Growth Hormone by Test and AGHD Group 22
Table 10. ITT and MAC: Agreement without categorization: Kendall's Tau-b 23
Table 11. ITT and MAC outcomes for All subjects 23
Table 12. ITT and MAC outcomes for groups B and C 23
Table 13. ITT and MAC: Agreement Levels for All Subjects 24
Table 14. ITT and MAC: Agreement levels for groups B and C 24
Table 15. ITT and MAC: Agreement levels at peak GH between 45 and 60 minutes (core study) 27
Table 16. Repeatability study (n=34): Agreement between ITT, first MAC, and second MAC tests 29
Table 17. ITT and MAC: Subgroup analyses by race 32
Table 18. ITT and MAC: Subgroup analysis by age 33
Table 19. ITT and MAC: Subgroup analysis by region 34
Table 20. ITT and MAC: Subgroup analysis by sex 34
Table 22. BMI and age subgroups 35

Reference ID: 4188461

LIST OF FIGURES

Figure 1. Study population 11
Figure 2. Scatter plot of ITT- and MAC-based GH levels for all study subjects....20
Figure 3. Scatter plot of ITT- and MAC-based GH levels by testing group....21
Figure 4. Boxplots of GH levels by AGHD groups....21
Figure 5. Kendall' s Tau-b: Agreement between MAC and ITT quantitative GH levels without use of cut points....22
Figure 6. Percent agreement between MAC and ITT tests in different subgroups....24
Figure 7. MAC: Distribution of GH by time and group 25
Figure 8. MAC: Means of GH levels by Time (core study) 26
Figure 9. MAC: Means of GH levels from repeated MAC test (repeatability test)....26
Figure 10. MAC: Group-specific Mean levels of GH by Time (core study) 26
Figure 11. Repeatability of MAC: scatter plot of GH values (core MAC versus repeated MAC, n=34) 27
Figure 12. Reproducibility: Changes in MAC-based GH from one test to another....28
Figure 13. Scatter plots identifying subjects with AGHD using post-hoc analysis 30
Figure 14. ITT and MAC: Separation between group A and group D....31
Figure 15. Subgroup analysis by region 33
Figure 16. GH by BMI 35
Figure 17 . BMI and GH distributions by group....36
Figure 18. ITT and MAC: Agreement by BMI....36
Figure 19. ITT and MAC: Box plot by age 37
Figure 20.Boxplots and q-q plots by country....42
Figure 21. Box plots and q-q plots by sex 42
Figure 23. Box plots and q-q plots by age (cut point age 25) 43

Reference ID: 4188461

1 EXECUTIVE SUMMARY

Aeterna Zentaris submitted an addendum for the NDA for macimorelin for the proposed indication of the diagnosis of adult growth hormone deficiency (AGHD) [redacted] (b) (4) . The goal of this review is to examine the performance characteristics and accuracy of macimorelin in the diagnosis of ADHD. This test is intended to be an alternative to the traditional insulin tolerance test (ITT), which is currently used in clinical practice though, the ITT test is associated with significant risk for the patient. Because of this issue, the ITT test is not commonly performed in the United States.

The submission is comprised of two Phase 3 trials (AEZS-130-047 and AEZS-130-052). Study AEZS-130-047 was previously submitted in November 2013. After the review cycle was completed, the Agency issued a Complete Response (CR) letter citing insufficient evidence to support the approval. Specifically, failure to reach the pre-specified primary goal for the area under the receiver operating characteristic curve (AUC), and questionable data quality were cited as deficiencies. Overall, the Agency was concerned with reproducibility of results of this study and recommended that the sponsor conduct a confirmatory trial. This study was reviewed by Dr. Lan Huang.

The focus of this review is the newly submitted study AEZS-130-052.

Performance characteristics of the macimorelin test: Based on the submitted data, the macimorelin test (MAC) does have skill in identification of subjects with AGHD. Although one of the co-primary endpoints was not met (achieving positive agreement with a lower confidence limit of 70% or more) the co-primary analyses did show a reasonable level of agreement between MAC and ITT in the study. The study provided the following results:

Positive agreement of MAC and ITT was 74.32% (95% CI: 62.84, 83.78);
Negative agreement of MAC and ITT was 93.94% (95% CI: 85.2, 98.32).

When the study population included only subjects with intermediate likelihood of AGHD, i.e., those who would be more likely to receive an AGHD diagnostic test in clinical practice, the agreement between the tests weakened:

Positive agreement of MAC and ITT became 61.11% (95% CI: 43.46, 76.86);
Negative agreement of MAC and ITT became 92.68% (95% CI: 80.08, 98.46).

My Recommendations:
Based on results from the pre-specified statistical analyses, the target performance characteristics for the macimorelin test identified in the primary hypothesis were not met. However, given the issue of adverse effects of the comparator ITT, and the fact that the test using macimorelin might have a better safety profile, and the fact that the study demonstrated evidence that this test does

provide a reasonable level of agreement with the ITT and has skill identifying subjects with AGHD, I recommend approval.
That being said, the label should only reflect the results obtained using the pre-specified cut point for diagnosis. (b) (4) [redacted] Definition of agreement metrics should also be added to the label.

Statistical issues and findings:

1. **Submitted study measures agreement, not accuracy.** A very important consideration in the interpretation of the outcome of this study is the following: given that there is no reference standard available for diagnosis of AGHD, the only way to evaluate a new test is to compare it with another test that is currently utilized in clinical practice for the same purpose. The comparator test used in this study, the ITT, likely does not have perfect accuracy. That is why this study should be deemed as an agreement study. The important point is that we are evaluating agreement between tests, and that agreement does not always mean accuracy.

2. **Intended use population.** Subjects with severe AGHD are not expected to be tested with the MAC or ITT test because their disease status would likely be determined by IGF-1 and prior medical history, thus leaving subjects with intermediate and low likelihood of AGHD as more representative of the intended use population for the MAC test. Therefore, the agreement estimates from the analysis of all groups together likely provides an inflated estimate of the performance of MAC in detection of AGHD, and results in the intermediate-likelihood subgroups B and C are important.

3. **One of the co-primary endpoints was not met.** The benchmark targets for both co-primary endpoints were not met, as the confidence interval for the percent positive agreement did not rule out the pre-specified threshold of 70%. However, the estimated levels of negative and positive agreement between the MAC test and the ITT were reasonably high (94% and 74%, respectively). Given the known adverse effects associated with administration of the ITT and the absence of other tests for GHD in the US, I do not have a problem accepting as sufficient the levels of agreement that were achieved with the pre-specified cut point. The safety profile of the MAC test might outweigh the limited efficacy (i.e., diagnostic performance), given that some agreement with the ITT test was demonstrated. But I should be clear that agreement for the intended use population, subjects with moderate or low likelihood of AGHD, is moderate. The study has several flaws summarized here with implications on conclusions and labeling recommendations.

4. **Secondary endpoints provided biased (overly optimistic) estimates** of sensitivity and specificity. These estimates only included subjects on the extremes (subjects with severe AGHD and healthy subjects). Therefore, they do not directly address important questions about diagnostic performance in the intended use population[2,3]. (b) (4) [redacted]. These types of evaluations are also not recommended for device clearance by CDRH[1].

5. **Healthy subjects came only from one center in one country**. Such selection of controls could limit the generalizability of the study results. Furthermore, subjects in the control group were homogeneous with respect to other baseline characteristics: all of them were white, only two subjects had BMI above 30 kg/m^2 and there were no subjects with BMI<35kg/m^2. Also, the small sample size and limited diversity of subjects made some of the subgroup analysis impossible to perform and the results of other subgroup analyses unreliable.

6. **Subgroup analyses of subjects with high BMI.** On multiple occasions (during type A and type C meetings conducted after study 47 received a complete response and prior to submission of study 52), FDA requested that the sponsor provide sufficient data (and potentially select different cut points based on BMI) to evaluate the MAC test when conducted in subjects with high BMI. This study did not include many subjects with high BMI and specifically, none of the subjects in the control group had high BMI above 35 kg/m^2. This issue could be mitigated by excluding subjects with high BMI from the intended use population in the labeling.

7. **Post-hoc analyses.** Because one of the co-primary endpoints was not met, the sponsor conducted a post-hoc analysis considering alternative cut points to the one that was pre-specified. For the original pre-specified analysis, the cut point for the MAC test was chosen based on results of a previous Phase 3 trial (study 47). In this post-hoc analysis, the sponsor used a data-driven approach (b) (4) [redacted]. In my view, post-hoc analysis is very useful as a learning tool that would be helpful to prepare for a new confirmatory study. The results of the post-hoc analyses could be an artifact of chance and could potentially overestimate the true performance of the test. To mitigate this issue, the outcomes of these post-hoc analyses should be validated on a separate independent dataset.

2 INTRODUCTION

2.1 Overview

A brief description of the drug indication and history of the submission is presented below.

2.1.1 Indication

Macimorelin is an orally available peptidomimetic tripeptide with growth hormone secretagogue activity similar to ghrelin. The proposed indication for this drug is diagnosis of adult growth hormone deficiency (AGHD) (b) (4) This drug is not designed for treatment, it is only intended to be used for the purpose of diagnosis.

2.1.2 History of Drug Development

Macimorelin was first developed under the sponsorship of Ardana Bioscience and then Aeterna Zentaris. The first study evaluating the diagnostic efficacy of macimorelin was started by Ardana as Study ARD-0705-001 and was renumbered AEZS-130-047 (study 47) after the transfer of the sponsorship to Aeterna Zentaris (AEZS).

The initial submission of NDA 205598 was on November 4, 2013. The efficacy part of this submission was focused on study 47. A Special Protocol Assessment (SPA) agreement for this study was reached on December 17, 2010. On November 5, 2014, FDA issued a Complete Response (CR) letter, citing statistical issues in study 47. Those issues were that the lower bound of the confidence interval for the area under the curve (AUC) of the receiver operator characteristic (ROC) curve was below the pre-specified threshold and that data quality was questionable because there was a lack of complete and verifiable source data for determining whether patients were accurately diagnosed with the target disease.

On March 6, 2015, the CR letter was followed up with a Type A meeting. During this meeting, the sponsor proposed a new study AEZS-130-052 (study 52) titled "Confirmatory validation of oral macimorelin as a GHST for the diagnosis of AGHD in comparison with the gold-standard insulin tolerance test (ITT)." At the Type A meeting, FDA requested that the sponsor recruit and conduct separate analyses in the higher BMI groups.

In January 2017, it was found that Study 52 did not meet one of two co-primary endpoints of the confirmatory efficacy analyses applying the predefined GH cut-off points for the macimorelin test and ITT. At a Type C meeting on March 29, 2017, FDA stated that analyses using different cut off points for the macimorelin test would be post hoc and would be interpreted in that light, requested additional subgroup analyses within the Group B and C patients who would be most likely to benefit from a new GH stimulation test and within obese and younger subjects, and stated that the adequacy of the performance of the new test would be a review issue.

2.1.3 Specific Studies Reviewed

This resubmission includes three additional studies (PK/PD trial (AEZS-130-054), QT trial (AEZS-130-055), and confirmatory trial (AEZS-130-052)). The focus of this review is the confirmatory trial, study 52. I did not review study 47, which was evaluated as part of the review of the original submission of NDA 205598. Additional details on the design and results of study 47 can be found in the Statistical Review by Dr. Lan Huang dated July 9, 2014.

Table 1. List of all studies included in analysis

Study	Phase and Design	Study Population
AEZS-130-052	**Phase 3 Open-label, randomized, multicenter, multinational, 2-way crossover study**	**Study population consisted of four groups:** **Group A: High likelihood of growth hormone deficiency (n=38)** **Group B: Intermediate likelihood of GHD (n=37):** **Group C: Low likelihood of GHD (n=40):** **Group D: Healthy control (n=25):**

2.2 Data Sources

This submission is in electronic common technical document (eCTD) format. The submission is archived at the following link:\\CDSESUB1\evsprod\NDA205598\0031\m5\datasets\aezs-130-052\analysis\adam\datasets

Study datasets were provided as SAS XPORT transport files. The analysis datasets were joinable by unique identifier (SUBJID). The datasets were in good organization. The Define.pdf file was clear enough. My analysis on the primary and secondary efficacy endpoints gives approximately the same results as those reported in the clinical study report (CSR).

I derived from the submitted datasets all of the results presented in this review. I created all tables and figures in this review unless otherwise noted.

3 STATISTICAL EVALUATION

3.1 Data and Analysis Quality

The submission quality was found to be reasonable.

3.2 Evaluation of Efficacy

Macimorelin is intended to be used as a diagnostic drug, and not intended for therapy. Therefore, the efficacy in this context is the ability of this test to separate subjects with disease, i.e. AGHD, from subjects who do not have disease. Because there is no reference standard for establishing

Reference ID: 4188461

the presence or absence of AGHD available, macimorelin is compared with the insulin tolerance test (ITT). The ITT test is currently used in clinical practice as one of the diagnostic tests for AGHD. Also, to show that a diagnostic test works, repeatability of the test needs to be confirmed.

3.2.1 Study Design and Endpoints

Study 52 was an open-label, randomized, single dose, 2-way crossover trial to examine performance characteristics of the macimorelin growth hormone stimulation test ('MAC') when compared with the insulin tolerance test in detection subjects with AGHD.

The study was conducted in 30 sites in 9 countries; 25 sites in Europe (Austria, Germany, Spain, France, Italy, Poland, Serbia, and UK) and 5 sites in the USA.

Study subjects were assigned to groups of descending likelihood of having AGHD. The groups were defined as follows:

Group A: High likelihood of GHD:

- Structural hypothalamic or pituitary lesions and low IGF-1, and/or
- Three or more pituitary hormone deficiencies (PHD) and low IGF-1, or
- Childhood onset GHD with structural lesions and low IGF-1

Group B: Intermediate likelihood of GHD:

- Eligible subjects not qualifying for either high or low likelihood (Group A/C)

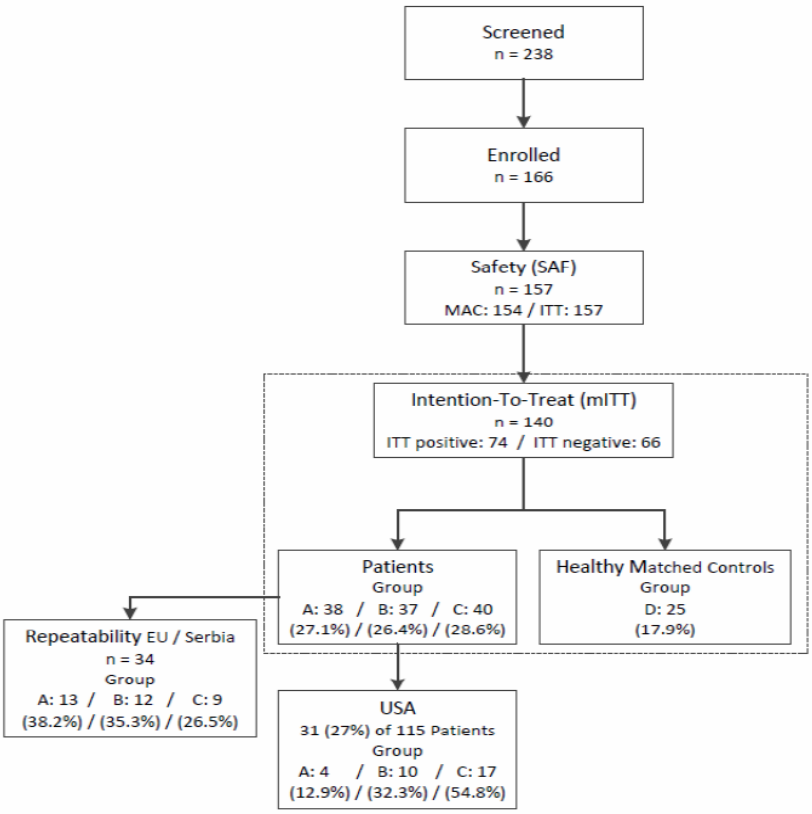
Group C: Low likelihood of GHD:

- One risk factor for GHD only, such as history of distant traumatic brain injury (TBI) or one PHD only with otherwise normal pituitary function, or
- Isolated idiopathic childhood onset GHD without additional pituitary deficits

Group D: Healthy control:

- Healthy subjects matching Group A subjects by sex, age, body mass index (BMI), and estrogen status (females only)

Figure 1. Study population



Out of 238 subjects screened, 166 subjects were enrolled into the study. A consort diagram showing subjects' disposition is presented in Figure 1.

After the core phase that involved MAC and ITT testing was completed, a small group of subjects (24.3%) was selected to repeat the MAC test in order to determine reproducibility of findings based on the MAC test.

For both the ITT and MAC tests, serum concentrations of growth hormone were measured at pre-defined time points (15, 30, 45, 60, 90) after GHST administration of macimorelin or insulin. According to the protocol, for the ITT, serum concentrations of growth hormone were also measured at 120 minutes after administration of insulin. Serum concentrations of GH for a given subject were provided by the central laboratory. Of note, the measurements performed at 120 minutes after test administration were mentioned in both the core study report (p. 25) and the clinical overview (p. 13), but the sponsor did not provide any analyses involving this time point and the measurements taken at 120 minutes after test administration were not present in the dataset provided to FDA.

Low levels of growth hormone indicate a growth hormone deficiency. Since the goal of both the MAC and ITT tests is to stimulate GH, peak levels of this marker indicate whether the patient has or does not have growth hormone deficiency.

The following cut points were pre-specified for the MAC and ITT tests based on the peak GH levels observed across the four measurement time points:

- MAC: GH: 2.8 ng/mL
- ITT: GH: 5.1 ng/mL,

where GH values below 2.8 for MAC and 5.1 for ITT identified subjects as having AGHD. The cut point selection for the MAC test was based on experience in a prior study (study 47). The cut point for ITT was selected based on the guideline for clinical practice.

The study cohort included all randomized subjects in whom both GHSTs were evaluable. The sequential order of the test was randomly determined for groups A, B, and C. Since subjects in group D were matched to subjects in group A, the order of testing in group D was determined by the order of testing selected for subjects in group A. The washout time between two tests was between 7 days and 1 month.

The analyses were performed using the modified Intent-to-Treat population, i.e. including only subjects who had both tests.

Of note, all subjects in the healthy control group (group D) were enrolled in one center in Poland. No controls were included from other sites. Repeatability of the MAC test was examined on selected subjects only in Serbia.

Intended use population:
In this submission, the sponsor does not clearly identify the intended use population. Following the text of this submission, the sponsor is intending the MAC test to replace ITT and therefore assuming that the intended use population for MAC test will be the same as intended use population for ITT test. The population of this trial includes subjects with severe disease identified through IGF-1 levels and prior history of pituitary lesions. In clinical practice, subjects with these characteristics will not be administered the ITT test because these parameters are clearly indicating growth hormone deficiency, and therefore, this group of subjects is not the population that will likely be getting the MAC test. Similarly, healthy subjects do not need testing. Therefore, the intended use population is more accurately represented by the subjects with intermediate (group B) and low likelihood (group C) of AGHD.

Primary endpoints:
Co-primary efficacy variables were ‘Percent Negative Agreement’ and ‘Percent Positive Agreement’ when using predefined cut-off points of both GHSTs. The benchmark for negative agreement was set to 75% and the benchmark for positive agreement was set to 70%. The goal for each of the co-primary endpoints was to show with high statistical confidence (based on the

lower bound of the Clopper-Pearson confidence interval) that the agreement was equal or higher than each of the prespecified benchmarks.

Secondary endpoints:
Secondary endpoints were 'Percent Overall Agreement' and performance characteristics, i.e. sensitivity and specificity of both GHSTs (MAC and ITT) when using predefined cut-off points of both GHSTs. No specific benchmark levels for those performance characteristics were pre-determined.

Sample size calculations:
The sponsor calculated sample size on the assumption that the MAC and ITT test results would be independent, which is not correct because this was a crossover trial and each subject experienced both tests during the trial. The planned sample size listed in the protocol was 110 subjects to have 55 ITT-positive and 55 ITT-negative subjects. The plan was to have at least 25% of subjects with high likelihood of having AGHD. The expected level of agreement (the sponsor did not provide a definition of what is agreement in this context) was to be at least 82% with the lower bound of the 1-sided exact confidence interval to be at least 69.3%.

3.2.2 Statistical Methodologies

Sponsor's approach:

There were two co-primary efficacy endpoints in this study: percent negative agreement and percent positive agreement, using the ITT as the comparator. For these analyses, the sponsor utilized the approach described in the "Statistical Guidance on Reporting Results from Studies Evaluation Diagnostic Devices" [1]. There is no reference standard (for example, pathology result) available for the diagnosis of AGHD, i.e. there is no specific test that can either confirm or deny presence of AGHD with perfect accuracy. In this scenario, performance characteristics of the new test (MAC) are evaluated based on comparisons to ITT, which is often used to diagnose AGHD in clinical practice but cannot be considered a true a reference standard as it cannot identify AGHD 100% correctly. In this setting, the performance of the new test can be based on the level of agreement between the new test and the comparator test.

As per the CDRH guidance for evaluation of agreement between two tests in circumstances of comparison to a non-standard, the performance characteristics of the MAC test were defined in the following manner:

Table 2. Definition of agreement

		Non-reference Standard (ITT)	
Macimorelin Test		+	–
	+	A	B
	–	C	D
Total		A+C	B+D

Percent Positive Agreement [%]=100% x A/(A+C)
Percent Negative Agreement [%]=100% x D/(B+D)
Percent Overall Agreement [%]=100% x (A+D)/(A+B+C+D)

The sponsor provided the two-sided 95% Clopper-Pearson confidence intervals for all calculated measures of agreement.

For the secondary endpoints, the sponsor compared subjects with high probability of AGHD (group A) with healthy subjects from group D. The sponsor-defined sensitivity (probability of positive test given presence of disease) and specificity (probability of negative test given absence of disease) were calculated based on the assumption that subjects in group A were "true" disease subjects and subjects from group D were "true" non-diseased subjects. Also, the sponsor conducted ROC analyses using only subjects from group A and D. For all these secondary analyses, subjects from groups C and B were excluded.

Both primary and secondary endpoints were examined in the subjects who underwent both the MAC and ITT tests.

Repeatability of the MAC test was examined on a subset of subjects. All subjects in the repeatability study were from one center in Serbia. For this evaluation, the subjects underwent a second MAC test. The difference between GH levels from both MAC tests was calculated.

FDA approach:

To thoroughly examine utility and external validity of a diagnostic test, it is important to address the following questions:

1. Who should be tested, i.e. what is the intended use population,
2. How accurate is the test (performance characteristics),
3. When should the test be conducted (timing of measurements), and
4. How reproducible are the test outcomes?

Intended use population
Regarding the intended use population, subjects with the highest probability of AGHD will most likely be identified by IGF-1 and prior medical history (as defined in selection criteria). Because

Reference ID: 4188461

these subjects could be identified without administration of an AGHD test, most likely they will not be receiving the test in clinical practice and therefore will not be a part of the intended use population.

Subjects from group D will most likely not be tested because they are healthy and therefore will not be a part of the intended use population. That would leave subjects with intermediate likelihood of AGHD, groups B and C, in the intended use population. Because of this, I conducted a separate subgroup analysis involving only groups B and C.

Performance characteristics of MAC
To address the accuracy of the MAC test, evaluation of concordance between MAC and the currently standard diagnostic test (ITT) for AGHD is appropriate. I agree with the definition of primary outcomes, i.e. use of percent positive and negative agreement to compare the new test to a non-reference standard.

In contrast to the definition of primary endpoints and definition of intended use population, the analysis and definitions of secondary endpoints proposed by the sponsor are problematic. For all secondary analyses, the sponsor excluded subjects from groups B and C. Thus, subjects who will likely be candidates for the ITT test (patients who would not be identified as having AGHD by IGF-1 or other criteria listed as criteria to identify subjects with high probability of AGHD) were excluded from the sponsor's evaluation. Furthermore, the sponsor-defined "sensitivity" and "specificity" are not valid estimates of those characteristics for the test in the intended use population. An analysis that does not include the intended use population but rather includes only subjects on the extremes introduces spectrum bias[1, 2, 3] and leads to biased estimates of sensitivity and specificity (tending to overestimate the performance characteristics of the test).

The outcomes of both AGHD tests are binary (disease or no disease) but the underlying biomarker (GH level) is quantitative; therefore, the binary test result is driven by selection of a cut point for the GH level. Because selection of a specific cut point for a biomarker is a subjective choice, it is important to learn how biomarkers (in this case GH obtained from ITT and GH resulting from the MAC test) relate to each other. Examination of agreement without use of cut points can provide a better understanding of GH patterns obtained with both tests. To explore the relationship between GH values obtained using ITT and GH values obtained using MAC test, I utilized Kendall's tau-b[4, 5]. Kendall's tau-b (tau-b) is an agreement measure that examines correlation between observations. The advantage of using tau-b is its scale-neutrality, it is a rank-based measure and because it is rank-based, subjects do not have to be on the same scale. The goal of this analysis is to measure the strength of association between biomarkers obtained using both tests. Tau-b was calculated for the entire study cohort and for each of the study groups separately.

I also used scatter plots and box plots to illustrate the relationship between the biomarkers.

Identifying optimal time window for testing
For every diagnostic test, it is important to identify optimal conditions so that the test outcomes provide the most accurate information. In the case of ITT and MAC tests, the optimal point will

Reference ID: 4188461

likely be obtained at the time when the patient has the maximum GH level. Since GH is measured multiple times after the administration of the test, it is important to find the time window when this biomarker is likely at its maximum. For this analysis, I examined the distribution of the GH values by time of the test. I compared distributions of GH obtained in each time window to the distribution of GH at maximum and identified the time point when the mean of the maximum GH distribution and the mean of GH in the time window were close. To replicate my findings, I utilized the data obtained during the MAC repeatability test. Finally, I compared agreement metrics obtained using maximum values of GH with those obtained from the time window that was closest to the maximum.

Reproducibility of MAC test

In addition to the sponsor's analysis, I created an illustrative waterfall plot that shows changes in GH from the initial MAC test and the second MAC test.

3.2.3 Patient Disposition, Demographic and Baseline Characteristics

Demographics and baseline characteristics are presented in Table 3 through Table 8. The overall age range of subjects was between 18 to 66 years old. Subjects in groups A and B were older (mean age 42.4 and 40.6 years old) than subjects in groups C and D (mean age 34.8 and 39.2 years old).

Table 3. Baseline demographic characteristics: Age distribution (years)

Group	Label	N	Median	Minimum	Maximum	Mean	Std Dev	Coeff of Variation
All	Age	140	41	18	66	40.6	13.1	32.2
A	Age	38	43.5	20	66	42.4	14.9	35
B	Age	37	48	20	65	45.7	11.9	25.9
C	Age	40	34.5	18	57	34.8	10.6	30.5
D	Age	25	39	18	62	39.2	12.2	31.2

The race distribution across all test groups was very heterogeneous. Black or African American subjects were only present in groups B (1 subject) and C (2 subjects). White subjects comprised most of the cohort (86.43%). Group D (healthy subjects) consisted entirely of White subjects.

Table 4. Demographic table: Race distribution

	RACE	Number of subjects	Percent Of the group
All	Asian	4	2.9
	Black or African American	3	2.1
	Other	2	1.4
	Unknown	10	7.1
	White	121	86.4
A	Asian	1	2.6
	Unknown	4	10.5
	White	33	86.8
B	Asian	1	2.7
	Black or African American	1	2.7
	Other	1	2.7
	Unknown	3	8.1
	White	31	83.7
C	Asian	2	5
	Black or African American	2	5
	Other	1	2.5
	Unknown	3	7.5
	White	32	80
D	White	25	100

The majority of study participants were in the lower BMI group (<30 kg/m^2). The highest fraction of subjects in the lowest BMI group was in the group with highest likelihood of AGHD (group A had 71% of subjects in the lower BMI group). Although the subjects in group D were supposed to match subjects in group A, the age and BMI distributions did not match well. Almost all subjects in group D were in the lowest BMI group (92%). Also, there were no subjects with BMI between 35 and 40 kg/m^2 in that group. In the previous correspondence and from the notes from the Type A meeting, FDA clearly requested that the sponsor recruit and conduct separate analyses in the higher BMI groups. However, there were only 15 total subjects (11%) in the study with a BMI between 35 and 40 kg/m^2.

Table 5. Demographic Table: BMI distribution

	BMI Group	Number of subjects	Percent Of the group
All	< 30 kg/m^2	98	70
	30 ≤35 kg/m^2	27	19.3
	35 ≤ 40 kg/m^2	15	10.7
A	< 30 kg/m^2	27	71.1
	30 ≤35 kg/m^2	7	18.4
	35 ≤ 40 kg/m^2	4	10.5
B	< 30 kg/m^2	21	56.8
	30 ≤35 kg/m^2	10	27
	35 ≤ 40 kg/m^2	6	16.2
C	< 30 kg/m^2	27	67.5
	30 ≤35 kg/m^2	8	20
	35 ≤ 40 kg/m^2	5	12.5
D	< 30 kg/m^2	23	92
	30 ≤35 kg/m^2	2	8

Gender was roughly equally distributed among groups A, B, and D. Only group C had a large majority of male subjects (80%). The majority of the study subjects were from the EU (number of subjects by country is presented in Table 7). Most of the subjects from the US were in the intermediate likelihood of disease groups (B and C). Group D consisted entirely of subjects from Poland.

Table 6. Demographic Table: Sex distribution

	SEX	Number of subjects	Percent of the group
All	F	57	40.7
	M	83	59.3
A	F	16	42.1
	M	22	57.9
B	F	22	59.5
	M	15	40.5
C	F	8	20
	M	32	80
D	F	11	44
	M	14	56

Table 7. Demographic Table: Region

	REGION	Frequency	Percent
All	EU	109	77.9
	US	31	22.1
A	EU	34	89.5
	US	4	10.5
B	EU	27	73
	US	10	27
C	EU	23	57.5
	US	17	42.5
D	EU	25	100

Table 8. Demographic Table: Country

	All		A		B		C		D	
Country	Subjects	%	Subjects	%	Subjects	%	Subjects	%	Subjects	%
AT	7	5	4	10.5	3	8.1				
DE	18	12.9	7	18.4	11	29.7				
ES	7	5	6	15.8	1	2.7				
FR	12	8.6	3	8	5	13.5	4	10		
IT	1	0.7			1	2.7				
PL	43	30.7	2	5.3	3	8.1	13	32.5	25	100
RS	20	14.3	12	31.6	2	5.4	6	15		
UK	1	0.7			1	2.7				
US	31	22.1	4	10.5	10	27	17	42.5		

Country codes: AT=Australia, DE=Germany, ES=Spain, FR=France, IT=Italy, PL=Poland, RS=Serbia, UK=United Kingdom, US=Unites States

Missing Data

Discontinuation of a subject from study procedures prior to randomization was not considered as a withdrawal but as a "screen failure." Subjects who withdrew or dropped out before completing the crossover study with both GHSTs being evaluable were replaced. The following describes the extent of study completion:

Number of ITT tested subjects: 157

- 27 (17%) out of 157 subjects did not achieve hypoglycemia that is needed for an ITT to be evaluable and required the repetition of the test
- Among the 17 ITTs that were actually repeated, 4 tests (24%) did not provide an evaluable result even on the "second try"
- 143 subjects had an evaluable ITT of which 3 withdrew prior to MAC
- Thus, the number of completed and evaluable ITT tests was 140, and the missing fraction was 17/157 or 10.8%

Number of MAC tested subjects: 154

- 1 subject had apparent non-compliance or a dosing error, and the MAC test was repeated
- Other MAC tests did not require repetitions
- Thus, the number of completed and evaluable MAC tests was 153, and the missing fraction was 1/154 or 0.65%

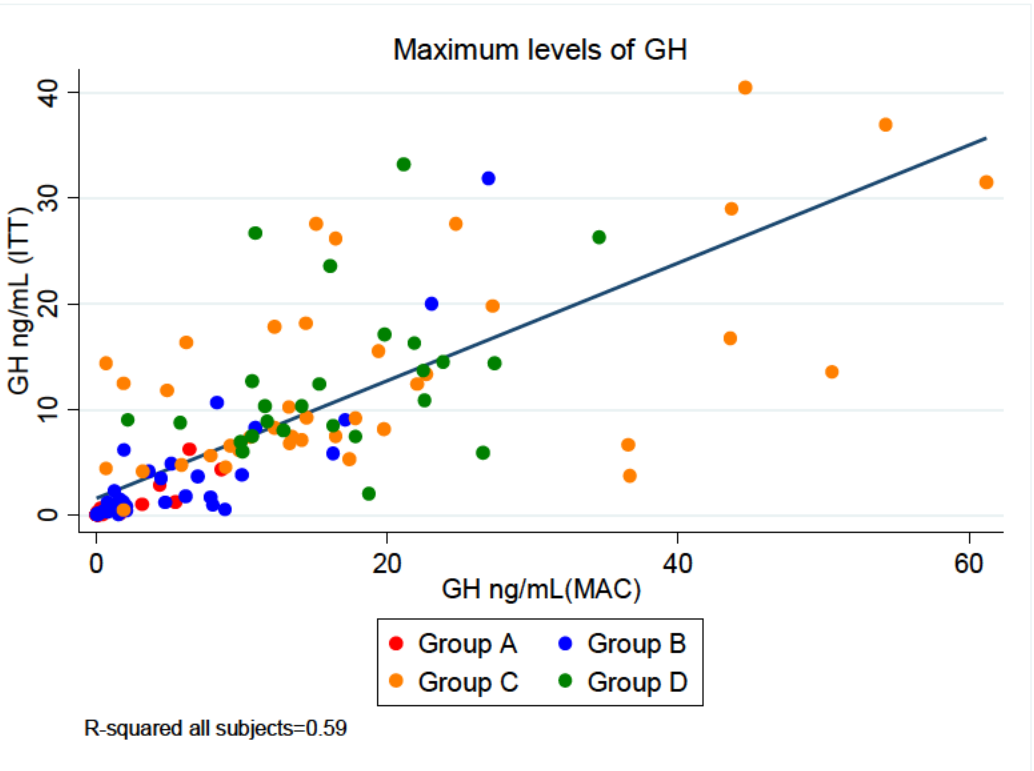
Overall, 140 out of the 157 enrolled subjects had both tests completed and evaluable (missing data in 17/157 or 10.8%). Because it was a crossover study, each subject underwent both tests. Therefore, the pre-specified primary analyses restricted to the subset of patients who completed both tests (the modified intention to treat population) still allows for a randomized comparison, and missing data for test comparisons was not as crucial as it would be in a parallel- group study. That being said, one reason for examining missing data in this setting is to examine the feasibility of obtaining test results. In this study, the ITT had most of the issues with feasibility

(17 out of 157 patients tested required repetition of the test, and 4 of those did not provide an evaluable result even on the second try). The MAC test showed almost no issues related to the feasibility of results (153 out of 154 subjects completed the MAC test without issues).

3.2.4 Results and Conclusions

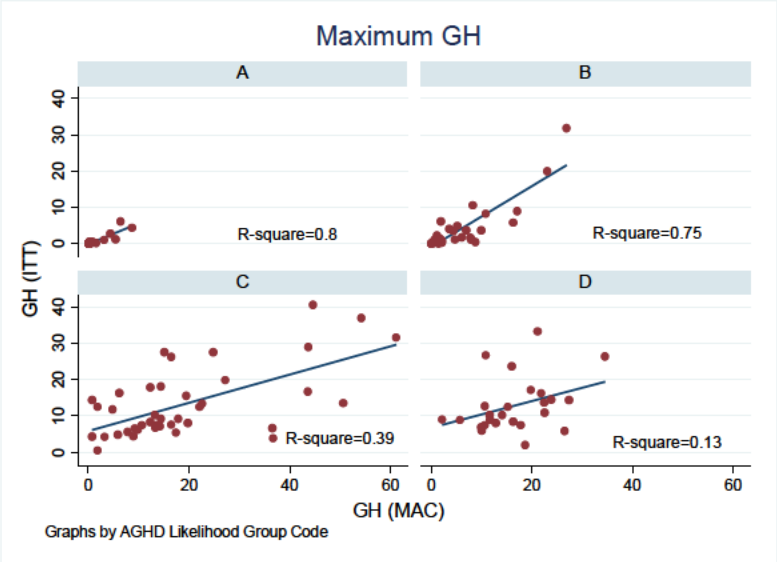
Unless otherwise noted, all of the following analyses and descriptive statistics were performed based on the maximum value of GH observed within each patient across the measurement time points. To illustrate the relationship between GH values obtained using the macimorelin test and ITT test, I created a scatter plot of raw data (Figure 2). The plot shows values obtained with ITT plotted against values with MAC. Each circle on the graph represents a value for one study participant. The purpose of this plot is to see how close the plotted circles align along a single straight line. Each group was plotted in a different color. Most of the higher levels of GH came from group C, although some of those subjects had high GH levels with MAC (e.g., >35ng/mL) but much lower values with the ITT test (e.g., <10 ng/mL). In addition, some of the subjects in group C had high GH from ITT-based levels and low GH from MAC-based levels.

Figure 2. Scatter plot of ITT- and MAC-based GH levels for all study subjects



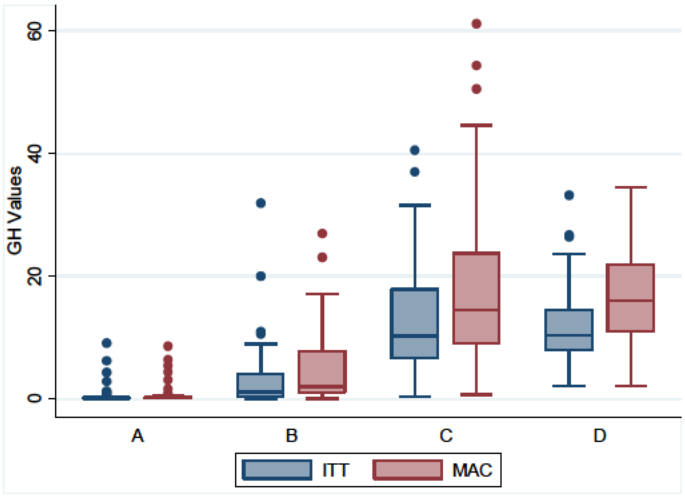
A more detailed illustration of GH distributions by group is presented in Figure 3. From this scatter plot, it becomes visible that all subjects in group A are clustered in the lower left corner indicating that GH levels were low for both ITT- and MAC-obtained results. The GH values for group C for both tests seem to have the largest variability.

Figure 3. Scatter plot of ITT- and MAC-based GH levels by testing group



A summary of distributional characteristics of GH by group is presented in the boxplot below (Figure 4). From this plot, it becomes clear that as the likelihood of AGHD goes up, the levels of GH increase regardless of what test is being used to obtain growth hormone values. The median values of GH levels in group C and D seem to be very similar. For the groups B, C, and D, the median value of MAC-obtained results is slightly higher than for the ITT-obtained results. A more detailed summary of numeric distributional characteristics of GH that was presented in boxplots is located in Table 9.

Figure 4. Boxplots of GH levels by AGHD groups



Legend: The boxes represent the 25th and 75th percentiles (interquartile range = IQR); solid lines within boxes, the median values; and vertical lines, 1.5 times the IQR; solid red and blue circles, the outliers exceeding 1.5 times the IQR.

Table 9. Levels of Growth Hormone by Test and AGHD Group

Group	Label	N	Median	Minimum	Maximum	Mean	Std Dev	Coeff of Variation
All	GH ITT test	140	4.4	0.1	40.5	7.2	8.9	123.8
A	GH ITT test	38	0.1	0.1	6.2	0.5	1.3	246.2
B	GH ITT test	37	1.2	0.1	31.9	3.5	6.2	176.4
C	GH ITT test	40	9.7	0.4	40.5	13.4	9.7	72.7
D	GH ITT test	25	10.3	2.1	33.2	12.9	7.5	58.4
All	GH MAC test	140	6	0.1	61.2	10.1	12.3	121.5
A	GH MAC test	38	0.1	0.1	8.6	0.9	2	213.7
B	GH MAC test	37	2	0.1	27	5.2	6.5	125.1
C	GH MAC test	40	14.5	0.7	61.2	19.2	15.5	80.8
D	GH MAC test	25	16.1	2.2	34.6	16.6	7.4	44.4

I explored agreement between the GH levels from the ITT and MAC tests without assumptions of any categorization. For this purpose, I used Kendall's Tau-b statistic. The results of my exploration are depicted in Figure 5 while detailed values are presented in Table 10. The highest level of agreement between tests was evident in group A. The overall agreement seems to mostly be driven by the results from this group. The lowest level of agreement and the largest uncertainty around the estimated level agreement was in group D. Overall, as the likelihood of AGHD increased, the level of agreement between ITT and MAC increased, suggesting that the tests perform differently for subjects without severe AGHD.

Figure 5. Kendall' s Tau-b: Agreement between MAC and ITT quantitative GH levels without use of cut points

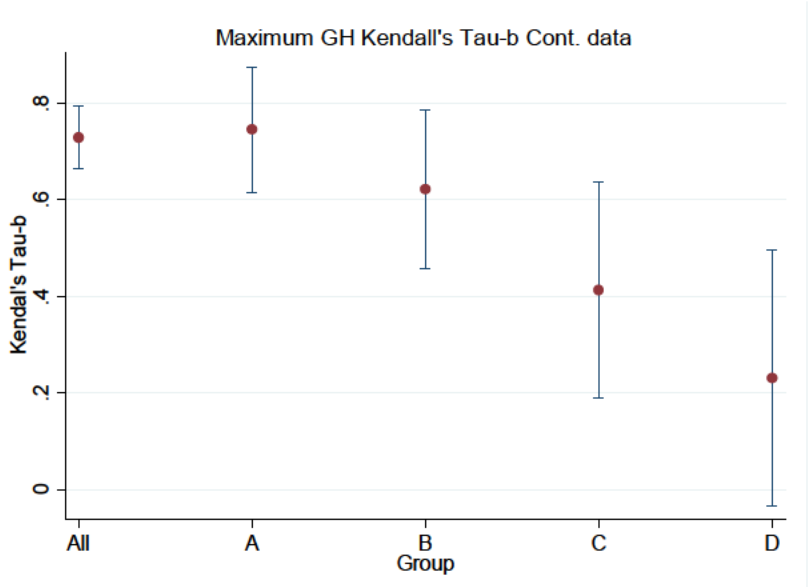


Table 10. ITT and MAC: Agreement without categorization: Kendall's Tau-b

Group	Tau-b	CI
All	0.73	(0.66, 0.79)
A	0.75	(0.61, 0.88)
B	0.62	(0.46, 0.79)
C	0.41	(0.19, 0.64)
D	0.23	(-0.03, 0.49)

Analysis of primary endpoints

The numbers of subjects categorized into test-positive (subjects who had GH values below pre-specified thresholds), and test-negative (subjects who had GH values at or above pre-specified thresholds) for each of the tests are presented in Table 11.
The results of analyses of the co-primary endpoints are presented in Table 13.
The estimated percent positive agreement between tests was 74% (95% confidence interval (CI): 63%, 84%). The estimated percent negative agreement was 94% (95% CI: 85%, 98%). The negative agreement benchmark (75%) was met, but the confidence interval for the positive agreement was clearly below the pre-specified benchmark of 70%. I also reevaluated the outcomes based on a more realistic representation of the intended use population consisting only of people from groups B and C (subjects with intermediate and low likelihood of AGHD). Based on this population, the numbers of subjects categorized into test-positive and test-negative for each of the tests are presented in Table 12. The results of agreement analyses, based on data from groups B and C, are presented in Table 14. After the test population was made to be close to the intended use population, the positive agreement went down from 74% to 61%. At the same time, the negative agreement did not change much (94% to 93%). The agreement measured with Kendall's Tau-b also went down from 0.73 to 0.62. Although the study was not powered to establish specifics levels of agreement for an analysis that includes only groups B and C, my findings suggest that the estimates of agreement using the entire cohort could be overestimating the true agreement between ITT and MAC that will be expected in the intended use population if the drug is approved to be used for testing.

Table 11. ITT and MAC outcomes for All subjects

		ITT test		
MAC test		+	-	
	+	55	4	59
	-	19	62	81
		74	66	

Table 12. ITT and MAC outcomes for groups B and C

		ITT test		
MAC test		+	-	
	+	22	3	25
	-	14	38	52
		36	41	

Table 13. ITT and MAC: Agreement Levels for All Subjects

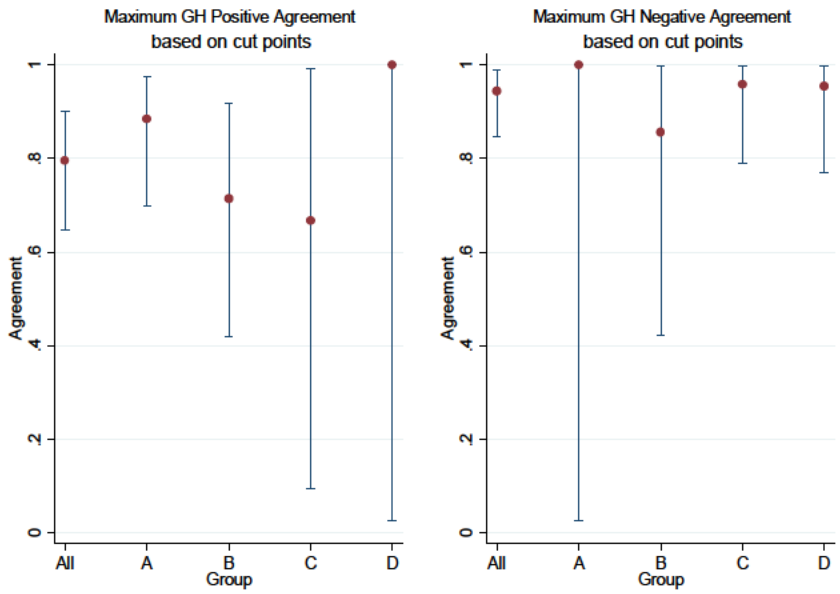
Agreement measure	Agreement level	95% CI
Kendall's Tau-b	0.73	(0.66, 0.79)
Positive Agreement	74.32	(62.84, 83.78)
Negative Agreement	93.94	(85.2, 98.32)

Table 14. ITT and MAC: Agreement levels for groups B and C

Agreement measure	Agreement level	95% CI
Kendall's Tau-b	0.62	(0.51, 0.73)
Positive Agreement	61.11	(43.46, 76.86)
Negative Agreement	92.68	(80.08, 98.46)

For the purposes of illustration, I performed analyses similar to the primary analyses for each group separately. The results are presented in Figure 6. The graph demonstrates that the highest levels of positive agreement were observed in group A. Groups B and C had much lower levels of positive agreement.

Figure 6. Percent agreement between MAC and ITT tests in different subgroups

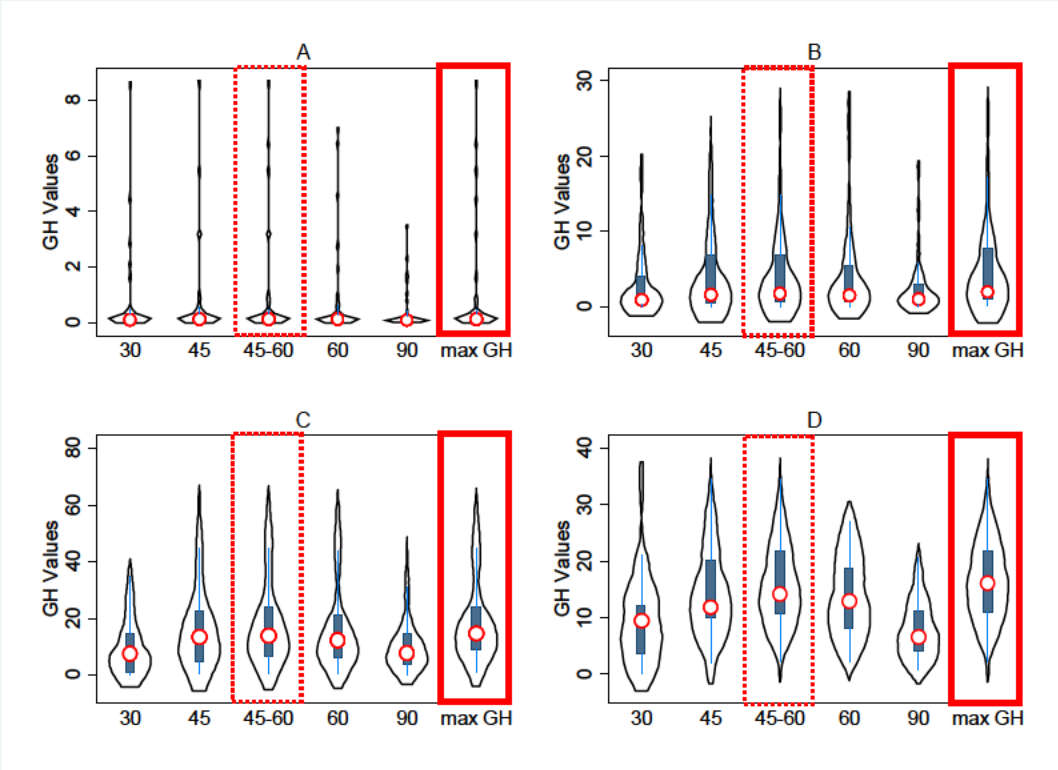


Selection of appropriate time point for testing

Distributions of GH by group and testing time are presented in Figure 7 below. The data presented in this graph show that the patterns of GH were very different among different groups. Specifically, as expected, GH values for subjects in group A tended to be lower than in other groups. GH distributions for subjects in group C tended to have greater variability than in all other groups. Because of the goal of the test, stimulation of growth hormone, the tests are likely to perform best utilizing GH values at the time point(s) when GH values within a patient reach their maximum. In the trial, subjects had four blood drawings: at 30, at 45, at 60, and at 90 minutes after administration of the drug. To explore whether such a frequent sampling is necessary, I compared distributions of GH at each blood draw, and from the violin plot, it seems

like distributions of the peak levels of GH across all 4 time points look similar to distributions of maximum levels of GH measured only at 45 and 60 minutes. The solid red rectangles in Figure 7 delineate distributions of maximum GH values across all 4 time points, while the red dotted rectangles delineate distributions of maximum levels of GH using only measurements at 45 and 60 minutes after drug administration.

Figure 7. MAC: Distribution of GH by time and group



Legend: Violin plots are visualizing distribution of the data and its probability density. Each plot is a combination of box plot and a density plot that is rotated and placed on each side to demonstrate the shape of the distribution that is being masked by visual limitations of boxplot. The boxes represent the 25th and 75th percentiles (interquartile range = IQR); red circles within boxes, the median values; and vertical lines, 1.5 times the IQR (outliers exceeding 1.5 times the IQR were removed from illustration).

I also examined the mean GH values at each of the blood draw time points. The results of my exploration are presented in Figure 8 below. For each time point, I plotted the mean and its 95% confidence interval. For the purposes of visual comparison, I plotted a horizontal red line which shows the estimate of the mean of the distribution of the maximum GH values (across the 4 time points). The mean value calculated for the maximum GH value measured in the time window between 45 and 60 minutes was closest to the mean of the maximum GH value across all 4 time points. This was also true based on results in the subgroup of subjects who underwent a second MAC test (Figure 9). The outcomes of the second MAC test in this subgroup were similar when analyzing each study group separately as shown in Figure 10.

Reference ID: 4188461

Figure 8. MAC: Means of GH levels by Time (core study)

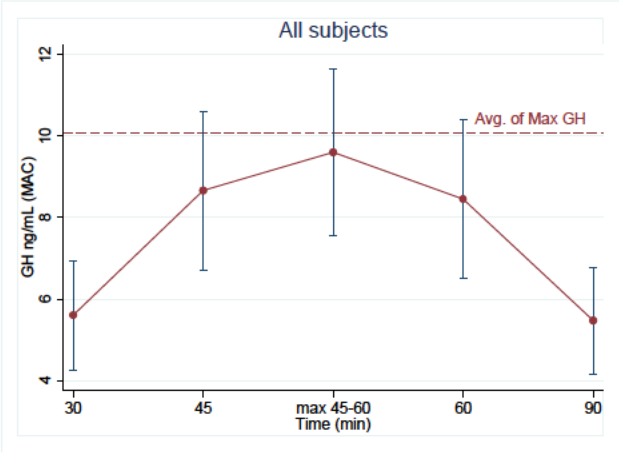


Figure 9. MAC: Means of GH levels from repeated MAC test (repeatability test)

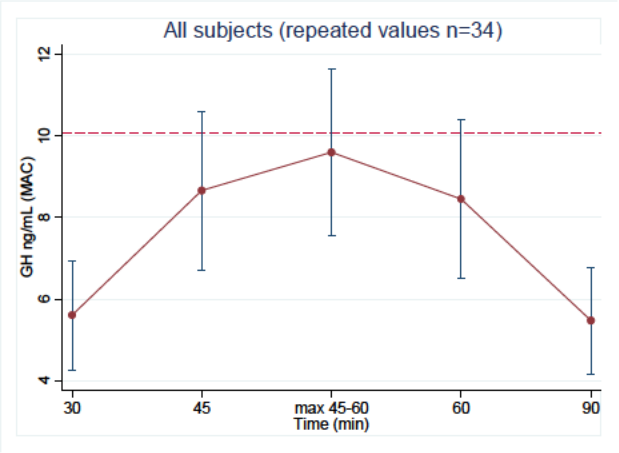
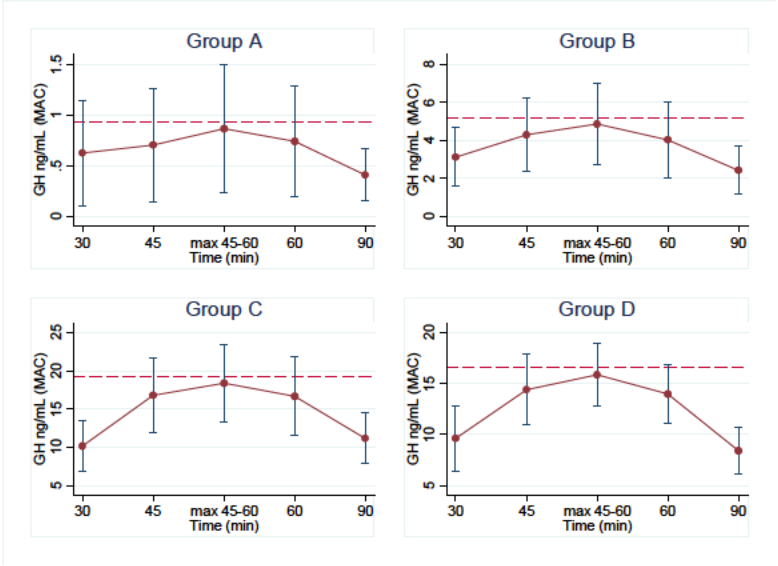


Figure 10. MAC: Group-specific Mean levels of GH by Time (core study)



Most importantly, I reexamined agreement between ITT and MAC when the MAC was performed using only the maximum of the 45 and 60-minute measurements (but the ITT was still carried out using the maximum across all 4 time points). The results of these findings are presented in Table 15. As it is shown, the results obtained with the MAC test using only the time window between 45 and 60 minutes are very close to the results obtained using maximum GH values across all 4 time points (positive and negative agreement of 76% and 94%, respectively, as compared to 74% and 94% using the pre-specified test). Thus, reducing the number of blood draws from 4 samples to 2 and obtaining them in the 45-60-minute time window might be a reasonable approach to consider, although these results are based on post hoc exploratory analyses.

Table 15. ITT and MAC: Agreement levels at peak GH between 45 and 60 minutes (core study)

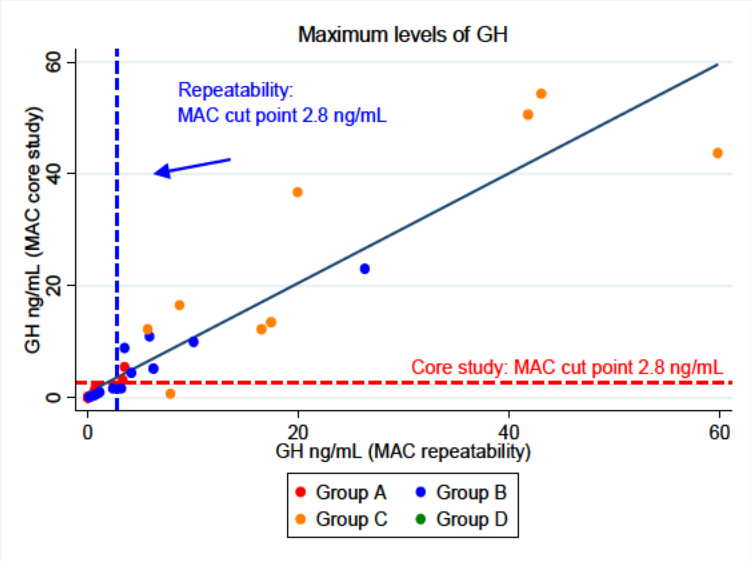
Agreement measure	Agreement level	CI
Kendall's Tau-b	0.70	(0.62, 0.77)
Positive Agreement	75.68	(64.31, 84.90)
Negative Agreement	93.94	(85.20, 98.32)

Reproducibility

To demonstrate that the GH values obtained using macimorelin are reproducible, the sponsor conducted repeatability testing. The sponsor administered a second MAC test to a subgroup of subjects from the core study (34 subjects (18 AGHD-positive and 16 AHGD-negative), all from Serbia). No healthy subject had a repeated MAC test. The goal was to see how well the results for the MAC test could be replicated.

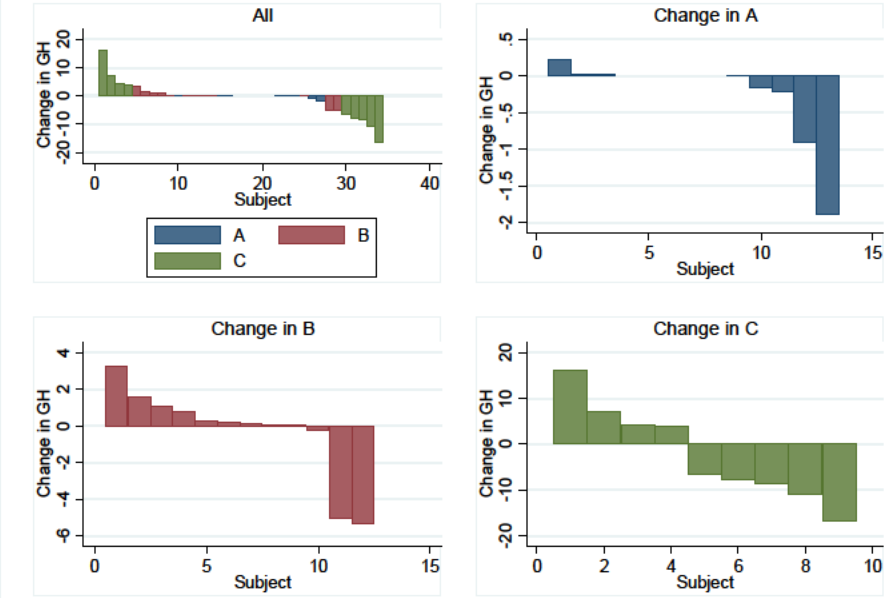
To illustrate the relationship between GH values obtained using the first MAC test and the second MAC test, I created a scatter plot of raw data. The plot shows values obtained with the repeated MAC plotted against values obtained with the first MAC in the core study. Each circle on the graph represents a value for one study participant. The purpose of this plot is to see how close the plotted circles align along a single straight line. The pre-specified cut points for both MAC tests are depicted by dashed lines. Like the scatter plot of ITT and MAC in the core study, each group was plotted in a different color. The GH values in groups A and B were closer to the straight line. The GH values for group C were further away from the straight line.

Figure 11. Repeatability of MAC: scatter plot of GH values (core MAC versus repeated MAC, n=34)



To illustrate changes in the levels of the GH, I created a group-specific waterfall plot that shows the differences in GH values between the initial MAC test and the second MAC test in Figure 12.

Figure 12. Reproducibility: Changes in MAC-based GH from one test to another



There were positive and negative changes in all groups, and some were quite large. The smallest changes were observed in group A and the largest changes were observed in group C. The changes in test-retest situations could be a result of intra-variability of GH within each subject.

As a result, only 2 subjects (5.9% of the repeatability cohort) changed their status from being classified as having AGHD by the first MAC test and classified as having no AGHD by the second MAC test. The disease status of other subjects remained unchanged.

The results of agreement analyses between ITT and each of the MAC tests (repeated and from the core study), and the agreement between MAC tests are presented in the Table 16 below. The agreement between both MAC tests was strong (0.86 for the maximum GH and 0.83 for the maximum GH measured in time window between 45 and 60 minutes). Interestingly, the agreement between the second MAC test and the ITT was slightly stronger than the agreement between the MAC test obtained during core study and the ITT (0.81 and 0.77 respectively).

Reference ID: 4188461

Table 16. Repeatability study (n=34): Agreement between ITT, first MAC, and second MAC tests

	ITT and MAC (core study)		ITT and MAC (repeated)		MAC (core study) and MAC (repeated)	
Time of blood draw	Kendall's Tau-b	95% CI	Kendall's Tau-b	95% CI	Kendall's Tau-b	95% CI
30 min	0.43	(0.2, 0.67)	0.37	(0.13, 0.61)	0.67	(0.5, 0.85)
45 min	0.62	(0.44, 0.79)	0.72	(0.59, 0.84)	0.72	(0.56, 0.89)
60 min	0.7	(0.56, 0.84)	0.81	(0.73, 0.89)	0.82	(0.73, 0.92)
90 min	0.72	(0.56, 0.88)	0.8	(0.68, 0.91)	0.84	(0.71, 0.97)
Max.	0.77	(0.63, 0.91)	0.81	(0.71, 0.91)	0.86	(0.77, 0.95)
Max. 45-60 min*	0.74*	(0.59, 0.88)*	0.81*	(0.71, 0.91) *	0.83**	(0.74, 0.93)**

*agreement between maximum (across all measurements) ITT-measured GH values and maximum (between 45 and 60 minutes) MAC-measured GH values
** agreement between maximum (between 45 and 60 minutes) MAC-measured GH values

Based on the small number of reclassified subjects and strong agreement between test results, it might be reasonable to consider that MAC test is reproducible, although these results are based only on the subgroup of subjects from one study site and did not include healthy subjects (subjects from group D).

Post-hoc and secondary endpoint analyses

Because one of the co-primary endpoints (the level of positive agreement) was not met, the applicant conducted a post-hoc analysis. For the original pre-specified analysis, the cut point for the MAC test of 2.8 ng/mL was chosen based on results of a previous Phase 3 trial and based on FDA recommendations to try to limit the likelihood of a false positive test. In this post-hoc analysis, the sponsor looked at all possible primary analysis outcomes for all possible cut points for the MAC test and selected a cut point yielding more favorable results, i.e., results in which the pre-specified benchmark for levels of agreement between the MAC test and ITT were satisfied.

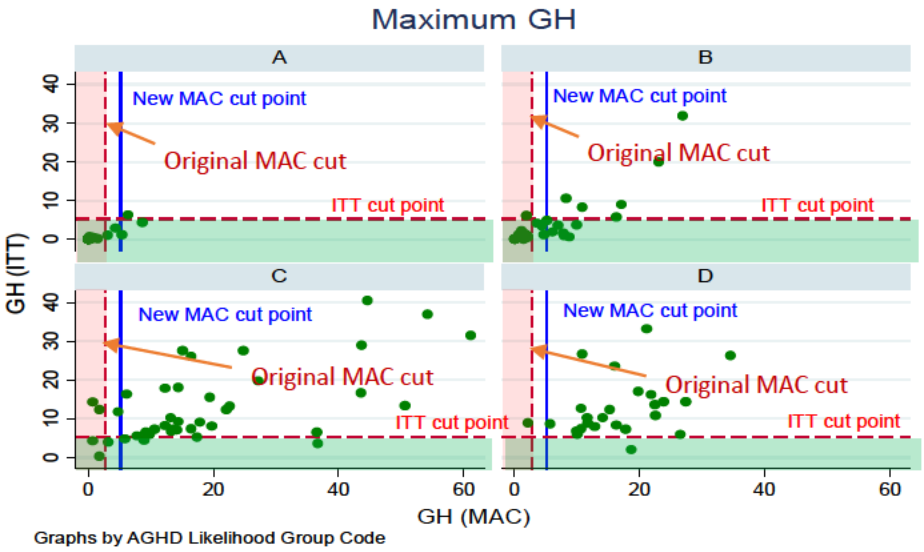
(b) (4)

A scatter plot showing the GH values based on the MAC and ITT tests, as well the pre-specified and post-hoc cut points is presented in Figure 13. In my view, post-hoc analysis is very useful as a learning tool that would be helpful to generate hypotheses to be tested in a new, independent confirmatory study. The results of post-hoc analyses, especially analyses that are clearly data-driven as were carried out here, could be an artifact of chance and are likely to overestimate the true level of agreement. The outcomes of post-hoc analyses should be validated on a separate independent dataset. (b) (4)

(b) (4)

However, the use of a higher cut point is likely to increase the likelihood that the MAC test will produce a false positive, and the study cannot reliably evaluate the degree to which this might occur (b) (4) .

Figure 13. Scatter plots identifying subjects with AGHD using post-hoc analysis



Analysis of secondary endpoints

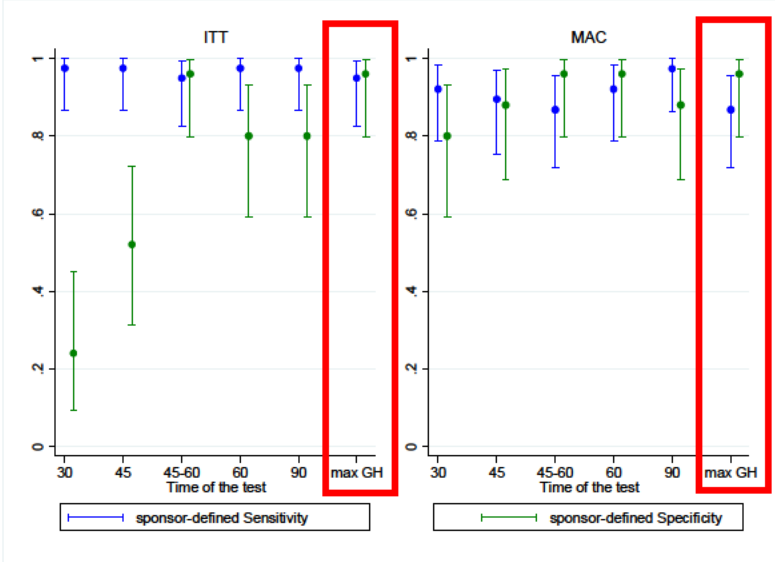
As I mentioned in the section "Performance characteristics of MAC", the definition of secondary endpoints proposed by the sponsor are problematic. Specifically, all secondary analyses excluded subjects from groups B and C. Thus, subjects who will likely best represent the intended use population of the test were excluded from evaluation.

I want to reiterate that sponsor-defined "sensitivity" and "specificity" are not true estimates of those characteristics for the intended use population. An analysis that does not include the intended use population is introducing spectrum bias. The proposed secondary endpoint evaluations do not address the question of the ability of the test to identify subjects in the intended use population; they only address the question of whether the test has any skill to identify subjects with AGHD who otherwise would have been identified through IGF-1 and other criteria, as well as the question of whether the test can indicate the absence of AGHD in healthy subjects (without likelihood of AGHD).

I am including a graph that shows sponsor-defined secondary endpoints, i.e., the probability that the ITT and MAC test correctly identifies subjects in groups A and D (Figure 14). This graph does not provide an accurate description of the performance of both tests for reasons described

above—these probabilities are not valid estimates of the sensitivity and specificity of the tests in the intended use population. The reason the "specificity" for the ITT test was low when only using GH values over the first 45 minutes is that blood samples for ITT were collected before the patient reached the state of hypoglycemia. This graph also shows that comparison between ITT and MAC tests by time, i.e. at time points when GH is not at its peak or close to the peak, is not meaningful because both tests have a different mechanism of action to stimulate rise of growth hormone.

Figure 14. ITT and MAC: Separation between group A and group D



3.3 Evaluation of Safety

Safety events were reviewed by Dr. William Lubas from Medical Division of Metabolism and Endocrinology Products. Readers are referred to Dr. Lubas' review for this section.

4 FINDINGS IN SPECIAL/SUBGROUP POPULATIONS

4.1 Gender, Race, Age, and Geographic Region

All subgroup analyses presented in this section were performed using the same approach as in the primary analysis. Because the study cohort consisted mostly of White subjects, the results for White subjects separately were similar to the overall cohort results (Table 17). The analyses in other subgroups might not provide reliable results because the numbers of subjects in those subgroups were small.

Table 17. ITT and MAC: Subgroup analyses by race

	Race subgroup	Agreement	Lower CL	Upper CL
Kendall's Tau-b	Asian	1	1	1
	Black or African American	1	1	1
	Other	1	1	1
	Unknown	0.64	0.42	0.86
	White	0.73	0.65	0.80
	All	0.73	0.66	0.79
Positive Agreement (%)	Asian	100	2.5	100
	Black or African American	100	2.5	100
	Other	100	2.5	100
	Unknown	66.67	29.93	92.51
	White	74.19	61.5	84.47
	All	74.32	62.84	83.78
Negative Agreement (%)	Asian	66.67	9.43	99.16
	Black or African American	100	15.81	100
	Other	100	2.5	100
	Unknown	100	2.5	100
	White	94.92	85.85	98.94
	All	93.94	85.2	98.32

The results of the subgroup analysis in subjects ages 65 years or more in Table 18 are not considered reliable because there were only 3 subjects in that age range. Therefore, only results for subjects younger than 65 have a meaningful interpretation. A more detailed analysis of the relationship between age and GH is presented in the next section.

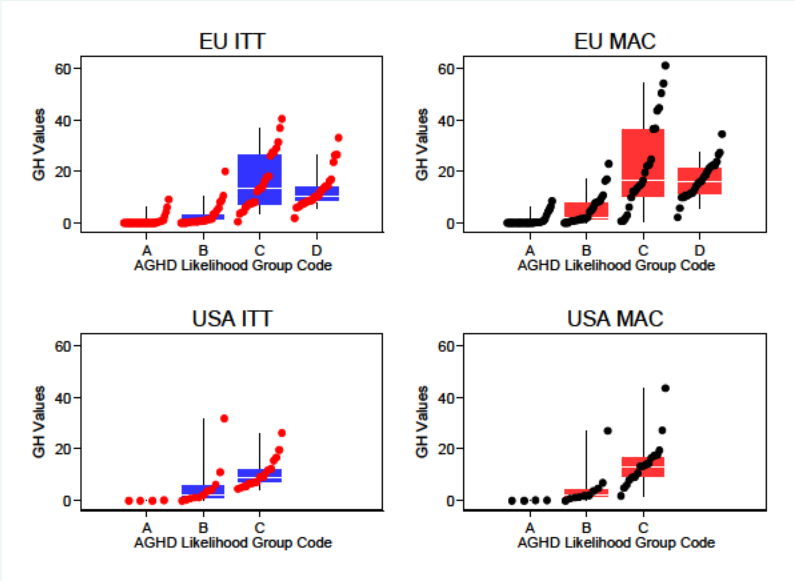
Reference ID: 4188461

Table 18. ITT and MAC: Subgroup analysis by age

	Age subgroup	Agreement	Lower CL	Upper CL
Kendall's Tau-b	<65	0.73	0.66	0.8
	≥65	0.33	-0.73	1
	ALL	0.73	0.66	0.79
Positive Agreement (%)	<65	76.06	64.46	85.39
	≥65	66.67	9.43	99.16
	ALL	74.32	62.84	83.78
Negative Agreement (%)	<65	93.94	85.2	98.32
	ALL	93.94	85.2	98.32

A graphical exploration of GH by region (US vs non-US) shows that the subjects from the US were mostly in the intermediate and low likelihood of AGHD groups (Figure 15). Those are the groups that showed the lowest levels of agreement between MAC and ITT. There were only four subjects with highest likelihood of disease (group A). Also, there were no subjects from the US in the control group because all controls came from one center in Poland.

Figure 15. Subgroup analysis by region



Legend: The graph consists of four panels: data for each region is presented for both MAC and ITT outcomes. Each panel shows boxplots of each group. The red circles that overlay the boxplots show an empirical quantile-quantile plot.

Potentially because most of the subjects from the US were in the intermediate and low risk group, the observed performance characteristics of the MAC test were worse than the

performance characteristics of this test in the EU (Table 19). A country-specific subgroup plot is presented in the Appendix (Figure 20).

Table 19. ITT and MAC: Subgroup analysis by region

	Region subgroup	Agreement	Lower CL	Upper CL
Kendall's Tau-b	EU	0.73	0.66	0.80
	US	0.69	0.52	0.86
	ALL	0.73	0.66	0.79
Positive Agreement (%)	EU	76.67	63.96	86.62
	US	64.29	35.14	87.24
	ALL	74.32	62.84	83.78
Negative Agreement (%)	EU	95.92	86.02	99.5
	US	88.24	63.56	98.54
	ALL	93.94	85.2	98.32

The subgroup analysis by sex showed no striking differences in test performance between men and women (Table 20).

Table 20. ITT and MAC: Subgroup analysis by sex

	Sex subgroup	Agreement	Lower CL	Upper CL
Kendall's Tau-b	F	0.67	0.54	0.79
	M	0.75	0.68	0.83
	ALL	0.73	0.66	0.79
Positive Agreement (%)	F	71.88	53.25	86.25
	M	76.19	60.55	87.95
	ALL	74.32	62.84	83.78
Negative Agreement (%)	F	84	63.92	95.46
	M	100	91.4	100
	ALL	93.94	85.2	98.32

4.2 Other Special/Subgroup Populations

Per to the study protocol, subjects in group D were supposed to be matched with subjects in group A by sex, age, BMI, and estrogen level for women. As shown in Table 21, there were no

Reference ID: 4188461

subjects in the 35-40 kg/m^2 BMI group among healthy participants. There were also no subjects in the entire study with BMI values above 40 kg/m^2. An important reason for assessing the MAC test in subjects (particularly, healthy subjects) with higher BMI is to ensure that false-positive readings for healthy subjects will be limited, because it is well known that subjects with higher BMI often have lower levels of GH without having AGHD[11]. Although, on many occasions (prior Type A and Type C meetings referenced in the history of drug development section), FDA specifically requested that the sponsor examine subjects with high BMI, the sponsor did not include many such subjects in the study.

Table 21. BMI and age subgroups

		AGHD likelihood group (n)				Total (n)
		A	B	C	D: Healthy	
BMI class	< 30 kg/m²	27	21	27	23	98
	30 - < 35 kg/m²	7	10	8	2	27
	35 - < 40 kg/m²	4	6	5	0	15
	Total	38	37	40	25	140
Age	18 - ≤ 25 years	7	2	10	5	24
	> 25 years	31	35	30	20	116
	Total	38	37	40	25	140

Source: Clinical Overview Table 4, p.14

To see the relationship between BMI and GH, please refer to Figure 16 and Figure 17 The level of agreement between MAC and ITT within subgroups defined by BMI is described in Figure 18.

Figure 16. GH by BMI

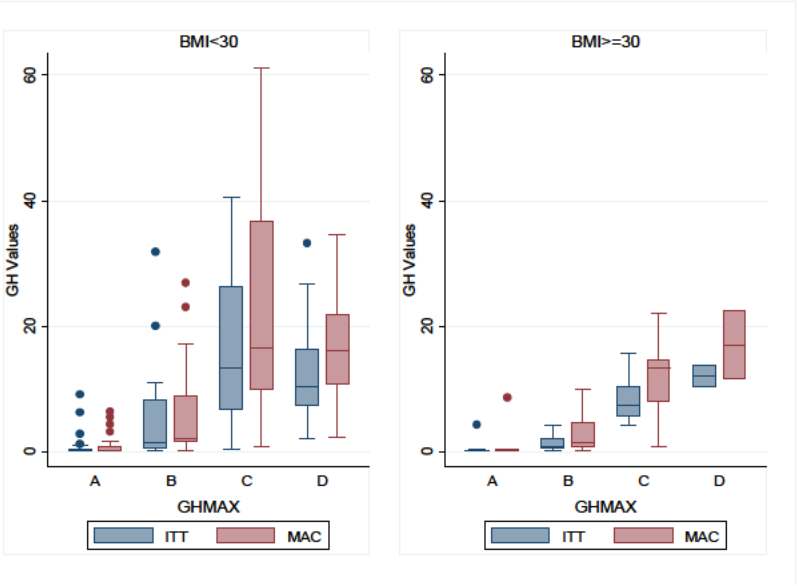


Figure 17 . BMI and GH distributions by group

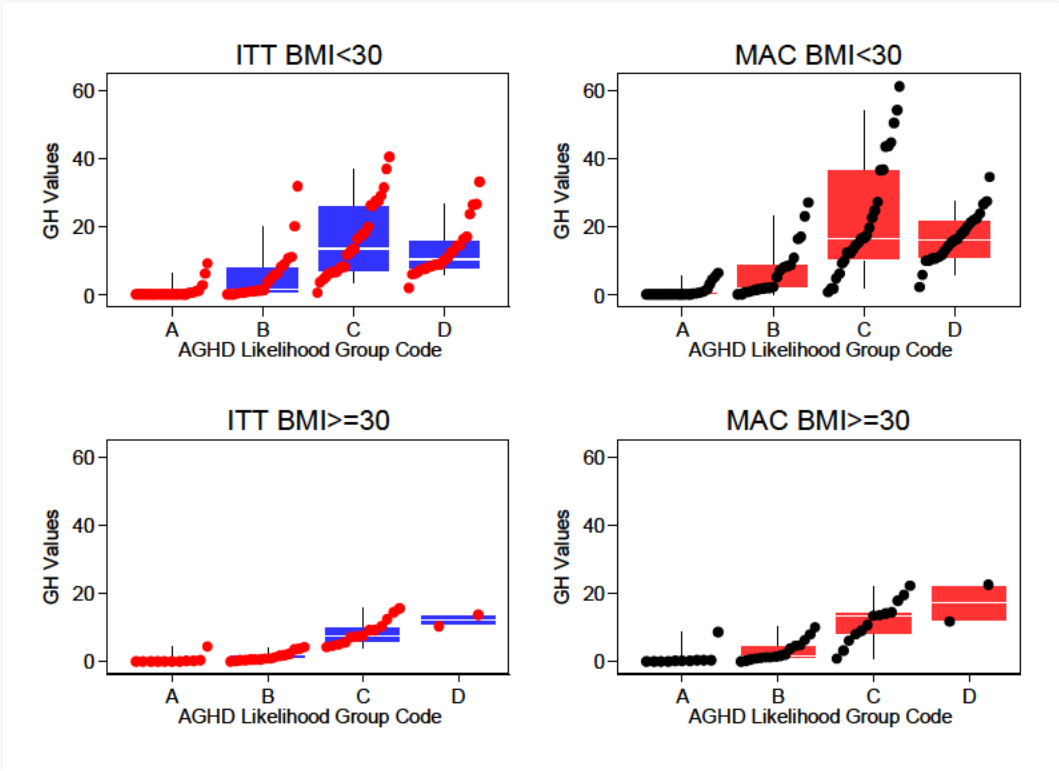
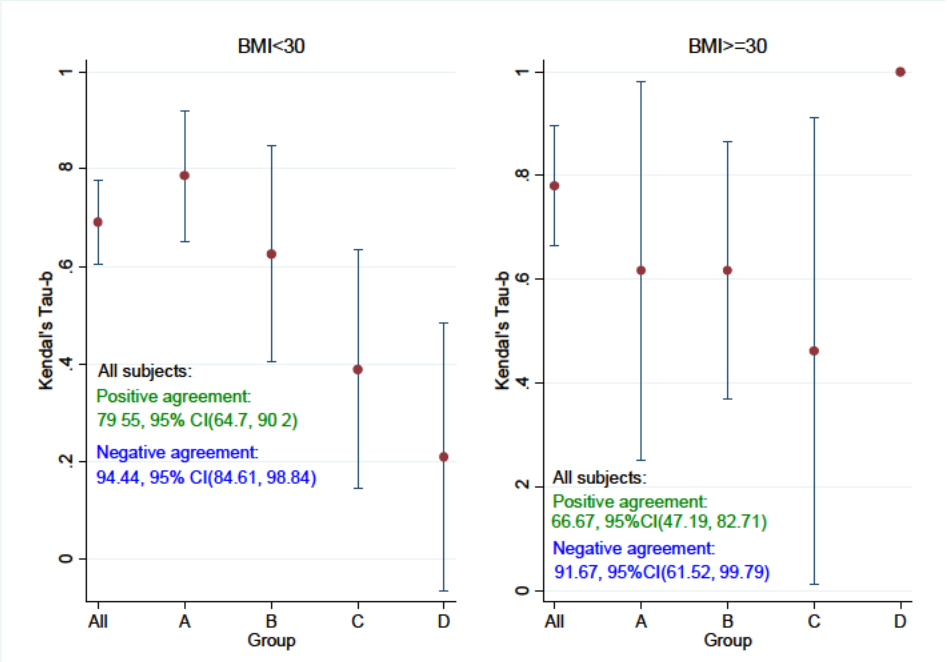
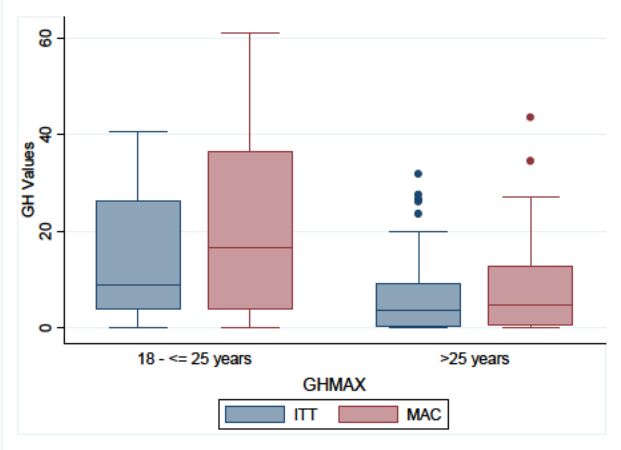


Figure 18. ITT and MAC: Agreement by BMI



It is also known that levels of growth hormone tend to decrease with increasing age. A box plot graph below shows that subjects age 25 and younger had higher GH levels than subjects in the older group. The reason for looking at subjects of age 25 or younger separately relates to the sponsor-selected subgroup for separate analyses (p. 46 core study report). Lower GH levels in older subjects could lead to reduced performance of diagnostic tests in this population. It is therefore potentially concerning that the sponsor did not include many subjects older than 65 years old in the study.

Figure 19. ITT and MAC: Box plot by age



5 SUMMARY AND CONCLUSIONS

5.1 Statistical Issues

1. **Submitted study measures agreement, not accuracy.** A very important consideration in the interpretation of the outcome of this study is the following: given that there is no reference standard available for diagnosis of AGHD, the only way to evaluate a new test is to compare it with another test that is currently utilized in clinical practice for the same purpose. The comparator test used in this study, the ITT, likely does not have perfect accuracy. That is why this study should be deemed as an agreement study. The important point is that we are evaluating agreement between tests, and that agreement does not always mean accuracy.

2. **Intended use population.** Subjects with severe AGHD are not expected to be tested with the MAC or ITT test because their disease status would likely be determined by IGF-1 and prior medical history, thus leaving subjects with intermediate and low likelihood of AGHD as more representative of the intended use population for the MAC test. Therefore, the agreement estimates from the analysis of all groups together likely

provides an inflated estimate of the performance of MAC in detection of AGHD, and results in the intermediate-likelihood subgroups B and C are important.

3. **One of the co-primary endpoints was not met.** The benchmark targets for both co-primary endpoints were not met, as the confidence interval for the percent positive agreement did not rule out the pre-specified threshold of 70%. However, the estimated levels of negative and positive agreement between the MAC test and the ITT were reasonably high (94% and 74%, respectively). Given the known adverse effects associated with administration of the ITT and the absence of other tests for GHD in the US, I do not have a problem accepting as sufficient the levels of agreement that were achieved with the pre-specified cut point. The safety profile of the MAC test might outweigh the limited efficacy (i.e., diagnostic performance), given that some agreement with the ITT test was demonstrated. But I should be clear that agreement for the intended use population, subjects with moderate or low likelihood of AGHD, is moderate.

4. **Secondary endpoints provided biased estimates** of sensitivity and specificity and do not directly address important questions about diagnostic performance in the intended use population. (b) (4) These types of evaluations are also not recommended for device clearance by CDRH.

5. **Subjects from the control group came only from one center in one country**. Such selection of controls could limit the generalizability of the study results. Furthermore, subjects in the control group were homogeneous with respect to other baseline characteristics: all of them were white with BMI<35kg/m^2. Also, the small sample size and limited diversity of subjects made some of the subgroup analysis impossible to perform and the results of other subgroup analyses unreliable.

6. **Subgroup analyses of subjects with high BMI.** On multiple occasions (during type A and type C meetings conducted after study 47 received a complete response and prior to submission of study 52), FDA requested that the sponsor provide sufficient data (and potentially select different cut points based on BMI) to evaluate the MAC test when conducted in subjects with high BMI. This study did not include many subjects with high BMI and specifically, none of the subjects in the control group had high BMI. This issue could be mitigated by excluding subjects with high BMI from the intended use population in the labeling.

7. **Post-hoc analyses.** Because one of the co-primary endpoints was not met, the sponsor conducted a post-hoc analysis considering alternative cut points to the one that was pre-specified. For the original pre-specified analysis, the cut point for the MAC test was chosen based on results of a previous Phase 3 trial (study 47). In this post-hoc analysis, the sponsor used a data-driven approach (b) (4)

(b) (4)

In my view, post-hoc analysis is very useful as a learning tool that would be helpful to prepare for a new confirmatory study. The results of the post-hoc analyses could be an artifact of chance and could potentially overestimate the true performance of the test. To mitigate this issue, the outcomes of these post-hoc analyses should be validated on a separate independent dataset.

Also, Dr. Lili Duan, Scientific Reviewer from CDRH (CDRH/OIR/DCTD), was a consultant for this review. Here is her comment on the validity of post-hoc analyses: "We do not allow post-hoc analysis to redefine test cut-points, except when there is prospective data to analyze and confirm the validity of the new cut-points. (b) (4) ."

5.2 Collective Evidence

My recommendation for approval is based on the fact, emphasized by the clinical team, that the existing alternative can be life-threatening and is therefore used sparingly in the United States. The statistical evidence for efficacy is marginal albeit, in my view, sufficient given this circumstance. The applicant's test did demonstrate agreement with the existing test although it did not satisfy the stringent prespecified primary endpoint.

The results of previously submitted study 47 did not provide sufficient evidence to support the claims proposed for macimorelin. Study 47 failed in the pre-specified primary ROC analysis. The ROC outcomes showed possible confounding of the test with race and study site. The sensitivity and specificity results were not consistent in different subject groups (previously vs. newly enrolled). Different post-hoc cut points were selected (not pre-specified). Those cut points were selected based on one dataset with 100 subjects. No validation of these outcomes was provided.

Based on the submitted data (study 52), the macimorelin test does have skill in identification of subjects with AGHD. Although one of the co-primary endpoints was not met (achieving positive agreement with a lower confidence limit of 70% or more), the co-primary analyses did show a reasonable level of agreement between MAC and ITT in the study. The study provided the following results using the prespecified (not post-hoc) cut point:
positive agreement of MAC and ITT was 74.32% (95% CI: 62.84, 83.78);
negative agreement of MAC and ITT was 93.94% (95% CI: 85.2, 98.32).

When the study population was reduced to include only subjects who would be more likely to receive an AGHD test, e.g., the intended use population, the agreement between the tests was weakened:
positive agreement of MAC and ITT became 61.11% (95% CI: 43.46, 76.86);
negative agreement of MAC and ITT became 92.68% (95% CI: 80.08, 98.46).

5.3 Conclusions and Recommendations

Currently, adult growth hormone deficiency is tested in clinical practice using the ITT test. Otherwise, clinicians rely on other factors, such as IGF-1 and prior medical history. The ITT introduces potential risks (because the patient must undergo an insulin-induced hypoglycemia), and is contraindicated in some patients. Because of these issues, the ITT is not commonly performed in the United States[8]. Although one of the co-primary endpoints was not met, given the unmet medical need for a safer diagnostic test, I would recommend approval if the safety profile of this drug meets FDA's requirements.

If another Phase 3 trial were conducted to evaluate the performance of MAC (or another AGHD diagnostic test), I recommend the following:

1. The study population should be representative of subjects in the intended use population
2. Control subjects should have diverse baseline characteristics and should not come from just one study site
3. Subjects from higher BMI groups should be a part of the study population
4. Similarly, because the sponsor would be seeking an approval in the United States, a sufficient number of US subjects should be included and the study should comply with current standards of clinical practice in the US
5. The results of post-hoc analyses (e.g., suggesting alternative cut points) should be confirmed using data from the new study.

5.4 Labeling Recommendations

I recommend approval with labeling based on the pre-specified cut point. (b) (4) [redacted] Definition of agreement metrics should be added to the label.

References:

1. FDA/CDRH. (2007). *Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests.* https://www.fda.gov/RegulatoryInformation/Guidances/ucm071148.htm#4

2. Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. New York: Oxford University Press.

3. Begg, C.G. (1987). Biases in the assessment of diagnostic tests. *Statistics in Medicine*, *6*, 411–423.

4. Kendall, M. (1938). *A New Measure of Rank Correlation*. Biometrika 30 (1-2): 81–89. doi:10.1093/biomet/30.1-2.81. JSTOR 2332226

5. Brown MB, Benedetti JK (1977) *Sampling Behavior of Tests for Correlation in Two-Way Contingency Tables.* JASA, 72, 309-315

6. Chanson P, Cailleux-Bounacer A, Kuhn JM, Weryha G, Chabre O, Borson-Chazot F, Dubois S, Vincent-Dejean C, Brue T, Fedou C, Bresson JL, Demolis P, Souberbielle JC (2010). *Comparative validation of the growth hormone-releasing hormone and arginine test for the diagnosis of adult growth hormone deficiency using a growth hormone assay conforming to recent international recommendations.* J Clin Endocrinol Metab. 2010 Aug;95(8):3684-92. doi: 10.1210/jc.2010-0295.

7. Molitch ME, Clemmons DR, Malozowski S, Merriam GR, Vance ML (2011); *Evaluation and treatment of adult growth hormone deficiency: an Endocrine Society clinical practice guideline*; Endocrine Society. J Clin Endocrinol Metab. 2011 Jun;96(6):1587-609. doi: 10.1210/jc.2011-0179. PMID: 21602453

8. Biller BM, Samuels MH, Zagar A, Cook DM, Arafah BM, Bonert V, Stavrou S, Kleinberg DL, Chipman JJ, Hartman ML (2002). *Sensitivity and specificity of six tests for the diagnosis of adult GH deficiency.* J Clin Endocrinol Metab. 2002 May;87(5):2067-79.

9. Stephanie M Chang, MD, MPH, David B Matchar, MD, Gerald W Smetana, MD, and Craig A Umscheid, MD, MS.(2012) *Methods Guide for Medical Test Reviews*; Agency for Healthcare Research and Quality (AHRQ US); 2012 Jun. Publication No.: 12-EHC017

10. Bartke, Andrzej et al. *Growth Hormone and Aging*. Journal of the American Aging Association 23.4 (2000): 219–225. PMC. Web. 24 Nov. 2017.

11. Bonert VS, Elashoff JD, Barnett P, Melmed S. *Body mass index determines evoked growth hormone (GH) responsiveness in normal healthy male subjects: diagnostic caveat for adult GH deficiency.* J Clin Endocrinol Metab. 2004 Jul;89(7):3397-401.

APPENDIX

Figure 20. Boxplots and q-q plots by country

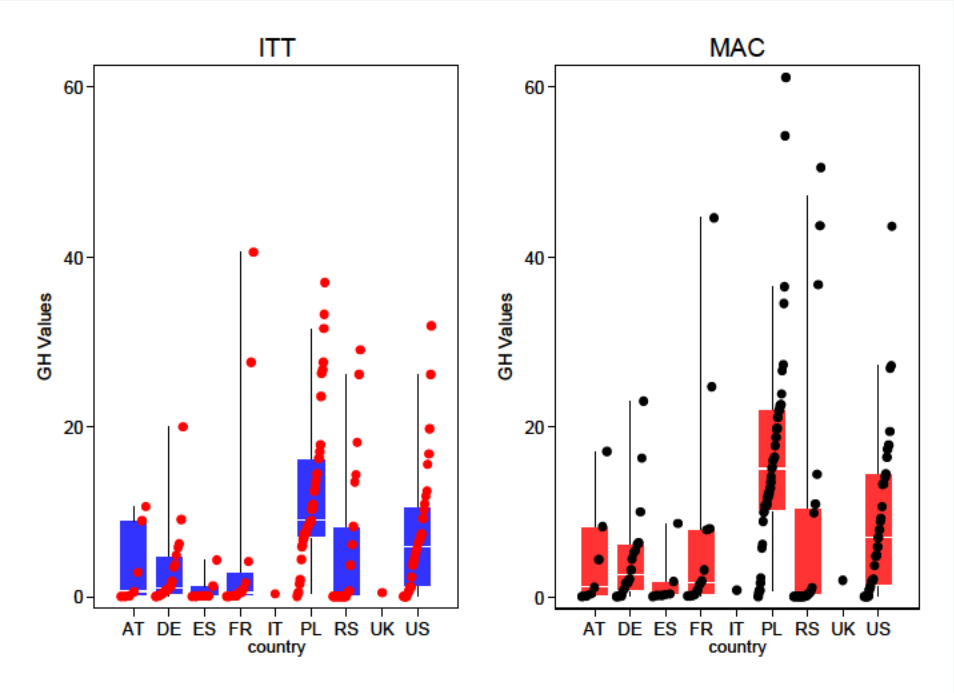
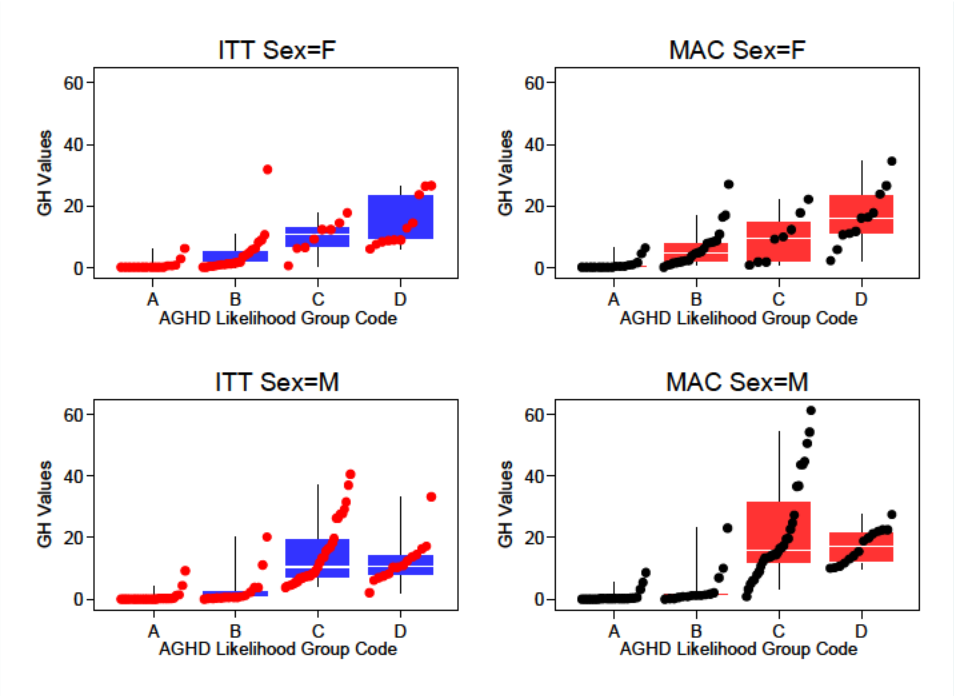
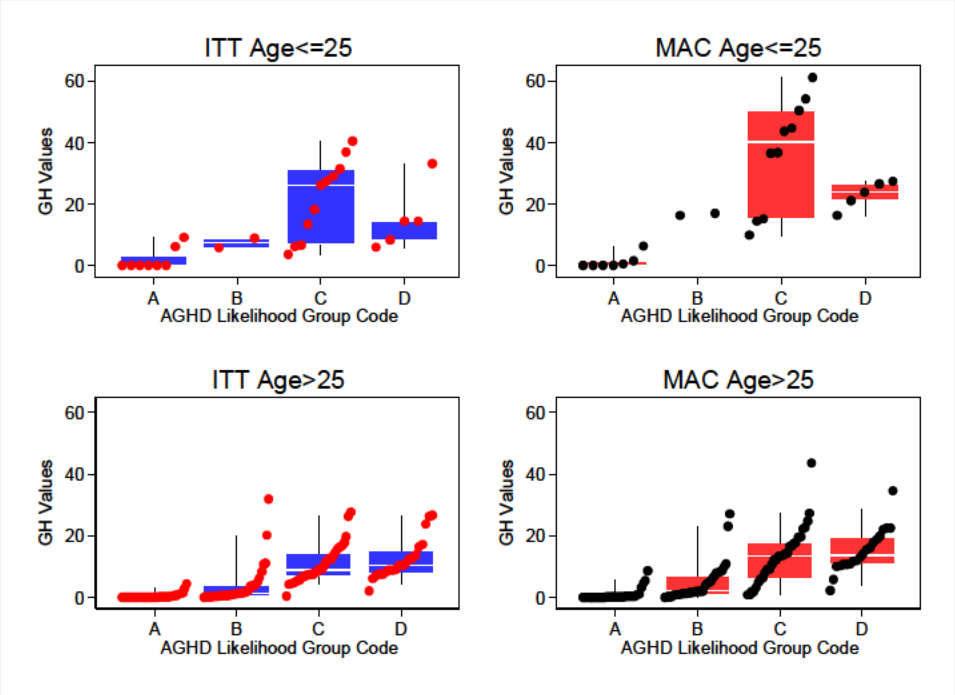


Figure 21. Box plots and q-q plots by sex



Reference ID: 4188461

Figure 22. Box plots and q-q plots by age (cut point age 25)



Reference ID: 4188461

This is a representation of an electronic record that was signed electronically and this page is the manifestation of the electronic signature.

/s/

ANNA E KETTERMANN
11/30/2017

GREGORY P LEVIN
11/30/2017

MARK D ROTHMANN
11/30/2017
I concur

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research
Office of Translational Sciences
Office of Biostatistics

STATISTICAL REVIEW AND EVALUATION

CLINICAL STUDIES

NDA/BLA #:	NDA 205-598/0000
Supplement #:	\\CDSESUB1\evsprod\NDA205598\205598.enx
Drug Name:	**Macimorelin Acetate Granules for Oral Solution**
Indication(s):	Diagnosis of Adult Growth Hormone Deficiency
Applicant:	Aeterna Zentaris GMBH, Germany
Date(s):	date submitted 11/04/2013
	PDUFA due date 11/05/2014
	Reviewer's due date 7/10/2014
Review Priority:	Standard
Biometrics Division:	Division of Biometrics V
Statistical Reviewer:	Lan Huang, Ph.D.
Concurring Reviewers:	Mark Rothmann (Team Leader)
Medical Division:	Division of Metabolic and Endocrine Products
Clinical Team:	William Lubas, M.D. (clinical reviewer)
	Roman Dragos (clinical division director)
Project Manager:	Abalade Adeolu

Keywords: Clinical Studies, NDA Review, diagnostic agent, ROC AUC, sensitivity and specificity

Table of Contents

1. EXECUTIVE SUMMARY 4

1.1 CONCLUSIONS AND RECOMMENDATIONS 4
1.2 BRIEF OVERVIEW OF CLINICAL STUDIES 4
1.3 STATISTICAL ISSUES AND CONCERNS 5

2. INTRODUCTION 6

2.1 OVERVIEW 6
2.1.1. CLASS AND INDICATION 6
2.1.2. HISTORY OF PROGRAM DEVELOPMENT 7
2.1.3. SPECIFIC STUDIES REVIEWED 7
2.1.4. MAJOR STATISTICAL ISSUES 8
2.2 DATA SOURCES 9

3. STATISTICAL EVALUATION 9

3.1 DATA AND ANALYSIS QUALITY 9
3.2 EVALUATION OF EFFICACY 9
3.2.1 Study Design and Endpoints 9
3.2.2 Statistical Methodologies 11
3.2.3 Patient Disposition, Demographic and Baseline Characteristics 13
3.2.4 Results and Conclusions 15
3.3 EVALUATION OF SAFETY 21

4. FINDINGS IN SPECIAL/SUBGROUP POPULATION 21

4.1 GENDER, RACE, AND AGE 21
4.2 OTHER SPECIAL/SUBGROUP POPULATIONS 21

5. SUMMARY AND CONCLUSIONS 23

5.1 STATISTICAL ISSUES 23
5.2 COLLECTIVE EVIDENCE 24
5.3 CONCLUSIONS AND RECOMMENDATIONS 25
5.4 LABELING RECOMMENDATIONS 25

Reference ID: 3539050

LIST OF TABLES

Table 1: Summary of pivotal clinical efficacy study AEZS-130-047 8
Table 2: Original study design 10
Table 3: Patient disposition (count) for subjects (all enrolled subjects) 13
Table 4: Demographics and Baseline characteristics for ITT/mITT population 14
Table 5: Distribution of subjects in mITT population by sites 15
Table 6: Summary of growth hormone concentration in AGHD patient and matched controls .. 16
Table 7: ROC analyses for mITT population 17
Table 8: Sensitivity, specificity, and percent of disagreement (in %) for mITT population with percent of disagreement <20% 18
Table 9: sensitivity and specificity for newly enrolled subjects and all subjects 19
Table 10: AUC for model with different drugs (test and comparator) 20
Table 11: sensitivity and specificity, and percent of disagreement for subjects with the comparator 20
Table 12: Sensitivity (sen) and specificity(spe) by subgroups (in %). 21
Table 13: Sensitivity (sen) and specificity (spe) point estimates by BMI and site (in %) using 2.7 ng/mL as the cut-point 23

Reference ID: 3539050

1. EXECUTIVE SUMMARY

1.1 Conclusions and Recommendations

The statistical results in terms of accuracy do not provide enough evidence to support the claims proposed in this NDA. The sponsor failed in the pre-specified primary ROC AUC analysis. The ROC AUC results may be biased because of possible confounding of the test with race and study site. The sensitivity and specificity results are not consistent in different subject groups (previously vs. newly enrolled). Different cut-points can be selected according to the rules (not pre-specified) and the cut-points are selected based on the single data with 100 subjects, which may not be reproducible in another study. We recommend CR with additional confirmatory study with pre-specified cut-point and intended patient population.

1.2 Brief Overview of Clinical Studies

The medical product under development is Macimorelin (AEZS-130), which is an orally-available peptidomimetic ghrelin receptor agonist with growth hormone secretagogue (GHS) activity. Macimorelin is intended for the diagnosis of adult growth hormone deficiency (AGHD) (b) (4)

This NDA submission included 5 studies in the clinical development program. Broglio et al, 2002, Study ARD-0705-003 and Study AEZS-130-IIT-1 included only healthy volunteers. Study H-27747 is an ongoing phase I study of macimorelin as a treatment of cachexia in cancer patients. Study AEZS-130-047 was selected for full statistical review and evaluation because it included both AGHD patients and matched controls. Macimorelin was administered to all ITT population in Study AEZS-130-047. The major objective of this study is to evaluate the diagnostic efficacy of macimorelin for adult growth hormone deficiency. The results of Study AEZS-130-047 are summarized below.

Study AEZS-130-047 provided a 99% Lower Bound for the one-sided confidence interval (CI) for the ROC AUC of .83 (less than the threshold of 0.85) for mITT population, thus the primary analyses (pre-specified in the SPA and agreed by both the sponsor and the agency) failed. For the per-protocol analysis (PPS) population defined by the sponsor after obtaining the data, the 99% Lower bound for the one-sided CI for the ROC AUC is 0.851 (>0.85).

52 subjects (42 AGHD patients + 10 matched controls) were enrolled previously under the original protocol and received both Macimorelin (AEZS-130) and L-ARG+GHRH (the comparator). 48 additional subjects (10 AGHD patients and 38 matched controls) were enrolled under the SPA agreed by the sponsor and the Agency and received Marcimorelin only because the comparator is not available during the study period.

The performance (see details in Table 9) for the new enrolled patients is worse than the previously enrolled subjects (80%<95% for sensitivity and 53%<90% for specificity, using

Reference ID: 3539050

the pre-specified cut-point 8.5 ng/mL). In addition, the specificity (30%) becomes worse if we only include the 10 newly enrolled matched controls for the 10 newly enrolled patients in the specificity analysis. The similar pattern is observed for another cut-point 2.7ng/mL determined using the data. The results are not consistent for subjects enrolled from different study portions.

With the cut-points (2-7 ng/.mL), the range of sensitivity and specificity values are from 80% to 90%, and percent of misclassification (disagreement between the test and the truth) ranges from 14% to 20% for the mITT population. A cut-point for the response variable (peak concentration following AEZS-130) of **2.7 ng / mL** produced a Sensitivity of 0.81 (95 CI: 0.68, 0.90) and a Specificity of 0.92 (95% CI: 0.80, 0.98) for mITT population.

The addition of a covariate (BMI) that separates the population into two groups – subjects with BMI < 30kg / M^2 versus subjects with BMI >= 30 kg / M^2 , followed by the ROC analysis performed independently on each of these groups, yielded the following results:
For mITT population, BMI-adjusted ROC AUC = 0.918. Lower bound of 99% CI is 0.845 by bootstrap method. Using the combined criteria of BMI with cut-point 30kg / M^2 and the peak concentration cut-point **6.8 ng/mL** in subjects with BMI<30 and **2.7 ng/mL** for subjects with BMI>=30, the sensitivity is 0.86 and specificity is 0.92 for the mITT population.

Subgroup analyses were conducted for the mITT population.
For the BMI < 30kg / M^2 group:
AUC = .892, lower bound of 99% CI is 0.975. Sensitivity is 77% and specificity is 85% with cut-point 6.8 ng/mL
For the BMI >= 30kg / M^2 group:
AUC = .939, lower bound of 99% CI is 0.851. Sensitivity is 83% and specificity is 96% with cut-point 2.7 ng/mL.

For the 52 subject received both Macimorelin (AEZS-130) and L-ARG+GHRH (the comparator), the AUC value (95% CI) is 0.986 (0.961, 1) and 0.929 (0.848, 1) for Macimorelin and L-ARG+GHRH, respectively. With the optimal cut-point **4.3 ng/mL** for Macimorelin, the sensitivity is 0.91 (95% CI: 0.77, 0.97), and the specificity is 1.00 (95% CI: 0.69, 1.00). With the optimal cut-point 7.4 ng/mL for L-ARG+GHRH, the sensitivity is 0.86 (95% CI: 0.71, 0.95), and the specificity is 0.90 (95% CI: 0.55, 1.00).

1.3 Statistical Issues and Concerns

Statistical significance was not achieved in the pre-planned primary analysis of the study. For the pre-planned primary analysis, the lower bound of the 99% confidence interval for the ROC AUC is <0.85, the pre-specified threshold.

There is a significant imbalance in Race between the AGHD group and the matched control group. Two of 52 AGHD patients (4%) and 18 out of 48 matched controls (38%) are Black or African American. Race may be a confounded factor for evaluating the diagnostic efficacy.

Reference ID: 3539050

There is a significant imbalance in study site between the AGHD group and the matched control group. For example, no AGHD patients were enrolled from Sites 11 and 17, and most of the matched controls were enrolled from Sites 09, 10, 11 and 17. Site may be a confounded factor for evaluating the diagnostic efficacy.

The determination of the optimal cut-point is a challenge. Different approaches may lead to different cut-points. In addition, the selected cut-point was based on the same data that was used to determine its sensitivity and specificity. Therefore, the estimates of sensitivity and specificity may tend to overstate the truth. It is not clear if the sensitivity and specificity obtained in Study AEZS-130-047 can be reproduced in another study.

For both sensitivity and specificity using cut-point 8.5 (pre-specified by the sponsor in the protocol) and 2.7 (optimal cut-point determined from the data by the sponsor), the performance for the newly enrolled patients is worse than the previously enrolled subjects. The results are not consistent (previously vs. newly enrolled).

It is not clear if the matched control group includes subjects belonging to the intended patient population, who would seek for diagnosis of possible AGHD and would benefit from the test. The performance (ROC AUC, sensitivity and specificity) of Macimorelin for the intended patient population is not known.

2. INTRODUCTION

2.1 Overview

2.1.1. Class and Indication

Macimorelin (**AEZS-130) is** an orally-available peptidomimetic ghrelin receptor agonist with growth hormone secretagogue (GHS) activity. Ghrelin potently stimulates growth hormone (GH) release. The GH-releasing effects of ghrelin are thought to be mediated by specific receptors mainly present at the pituitary and hypothalamic level. *In vitro*, utilizing membrane preparations containing the GHS-receptor derived from human hypothalamus and pituitary gland, as well as in various functional cell-based assays covering GHS receptors of humans, mice and rats, it was demonstrated that macimorelin shows binding potency to the human GHS receptor comparable to that of its natural ligand, ghrelin. *In vivo*, macimorelin is readily absorbed from the gastrointestinal tract and induces potent GH release.

The proposed **Indications and Usage statement** for Macimorelin (by the sponsor) is:

(b) (4)

- MACRILEN is indicated for the diagnosis of adult growth hormone deficiency (AGHD)

(b) (4)

(b) (4)

2.1.2. History of Program Development

Macimorelin has been developed as a diagnostic agent for Growth Hormone deficiency (GHD) in adults, first under the sponsorship of Ardana Bioscience and then Aeterna Zentaris.

At a pre-IND meeting with the FDA, Ardana (b) (4) FDA suggested (b) (4) to allow receiver operating characteristic (ROC) curves to be constructed and classification and regression tree (CART) analyses to be performed (for definitions and illustrations of these analyses). The Division recommended an alternative design based upon the publication entitled "Sensitivity and Specificity of Six Tests for the Diagnosis of Adult GH Deficiency (Biller et al, 2002). In terms of other protocol design questions, safety endpoints (including vital signs and electrocardiograms) were deemed adequate and 30-40 control subjects and 30-40 subjects with presumed GHD were considered sufficient. Ardana agreed to this alternative study design.

Because one of the two originally planned treatment arms (GHRH + L-ARG (arginine)) became unavailable during the conduct of the study and after a series of discussions and meetings with the US Agency, a revised protocol and request to review under the Special Protocol Assessment (SPA) procedure was submitted on 14 July, 2010. The revised protocol and SPA were agreed by the Agency and Aeterna Zentaris on 17 December, 2010. Subsequently the trial was conducted and completed successfully by Aeterna Zentaris. Because this study was conducted under an SPA, it forms the basis for approval of macimorelin acetate (granules for oral solution) as a diagnostic agent for AGHD in this NDA.

(b) (4) In a separate ongoing clinical trial, it is also being studied as a therapeutic agent in adults suffering from cancer cachexia.

2.1.3. Specific Studies Reviewed

Four clinical studies with macimorelin acetate have been completed and one study is underway. The five clinical studies are:

- Broglio et al, 2002: two young healthy male volunteers received IV macimorelin (1 µg/kg) and incremental doses of oral macimorelin (0.06, 0.125, 0.25 and 0.5 mg/kg).
- Study ARD-0705-003: the effects of a meal on the oral bioavailability of macimorelin 0.5 mg/kg were evaluated in healthy volunteers.
- Study AEZS-130-IIT-1: macimorelin was administered in increasing single oral doses (0.005 to 0.5 mg/kg) and intraduodenal doses (0.2 to 0.5 mg/kg) in healthy male volunteers.
- Study AEZS-130-047: macimorelin was administered as a single oral dose of 0.5 mg/kg in adult growth hormone deficiency (AGHD) patients and in matched controls. Control subjects were matched to AGHD patients based on age (± 5 years), gender, BMI (± 2 kg/m2) and estrogen supplement status (for women only).

- Study H-27747: An investigator-initiated double-blind Phase I study of macimorelin as a treatment of cachexia in cancer patients. As of July 1, 2013 in this ongoing study, patients have been enrolled only in the dose group of oral macimorelin 0.5 mg/kg or matching placebo, once a day for 7 days.

In this review, we only evaluate the results obtained from the pivotal study AEZS-130-047 with both AGHD patients and matched controls.

Table 1: Summary of pivotal clinical efficacy study AEZS-130-047

Phase and design	Primary objective	Study period	treatment	Subjects (No. (M/F), type)	Primary endpoint
Phase 3, multi-center (11 USA) open label	Diagnostic efficacy for adult growth hormone deficiency	07/12/2007 to 7/14/2011 with total of 100 subjects with macimorelin exposure	MacimorelinOral solution Single dose 0.5 mg/kg	52 (31/21) AGHD patients 48 (30/18) Matched controls	ROC AUC* of peak GH concentrations

*Area under the receiver operating characteristic curve

2.1.4. Major Statistical Issues

1) Statistical significance was not achieved in the pre-planned primary analysis of the study. For the pre-planned primary analysis, the lower bound of the 99% confidence interval for the ROC AUC is <0.85, the pre-specified threshold.

2) There is a significant imbalance in Race between the AGHD group and the matched control group. Two of 52 AGHD patients (4%) and 18 out of 48 matched controls (38%) are Black or African American. Race may be a confounded factor for evaluating the diagnostic efficacy.

3) There is a significant imbalance in study site between the AGHD group and the matched control group. For example, no AGHD patients were enrolled from Sites 11 and 17, and most of the matched controls were enrolled from Sites 09, 10, 11 and 17. Site may be a confounded factor for evaluating the diagnostic efficacy.

4) The determination of the optimal cut-point is a challenge. Different approaches may lead to different cut-points. In addition, the selected cut-point was based on the same data that was used to determine its sensitivity and specificity. Therefore, the estimates of sensitivity and specificity may tend to overstate the truth. It is not clear if the sensitivity and specificity obtained in Study AEZS-130-047 can be reproduced in another study.

5) For both sensitivity and specificity using cut-point 8.5 (pre-specified by the sponsor in the protocol) and 2.7 (optimal cut-point determined from the data by the sponsor), the

Reference ID: 3539050

performance for the newly enrolled patients is worse than the previously enrolled subjects. The results are not consistent (previously vs. newly enrolled).

6) It is not clear if the matched control group includes subjects belonging to the intended patient population, who would seek for diagnosis of possible AGHD and would benefit from the test. The performance (ROC AUC, sensitivity and specificity) of Macimorelin for the intended patient population is not known.

2.2 Data Sources

All materials reviewed including the applicant study reports, data sets and literature referenced are provided electronically, and the full electronic path of the documents are \\CDSESUB1\evsprod\NDA205598\205598.enx

The application study reports reviewed include Clinical overview, Summary of Clinical Efficacy, Summary of Clinical Safety in M2.

Data sets analyzed for study AEZS-130-047 (with data definition document) were located in M5\datasets\aezs-130-047\analysis\adam\datsets.

The data follow Analysis Data Model (ADAM) and Data Tabulation Model (SDTM) data format.

3. STATISTICAL EVALUATION

3.1 Data and Analysis Quality

It is possible to reproduce the primary analysis dataset, and in particular the primary endpoint, from the original data source. It is easy to find the related data. Moderate level of effort is needed to process the data.

3.2 Evaluation of Efficacy

3.2.1 Study Design and Endpoints

AEZS-130-047:

Primary Objective:
To determine the diagnostic efficacy of macimorelin (AEZS-130) based upon the area under the ROC curve (ROC AUC)

Secondary Objectives:
• To define cut-points for AEZS-130 based upon the sensitivity (ROC), specificity (ROC), and misclassification (CART) analyses
• Assessment of effect on IGF-I, if any

Reference ID: 3539050

• Evaluation of AEZS-130 safety profiles in terms of: Incidence of treatment-emergent AEs, Laboratory findings, Vital signs, ECGs at Tmax of AEZS-130 for the ~40 additional control subjects and 10 additional AGHD subjects to be entered under Amendments 4 and 5.

Study design:
This study was originally designed as a Phase II/III multi-center, randomized, open-label, cross-over trial in case-control subjects. The following graphic depicts the original design:

Table 2: Original study design

		Period 1: Day 1	Period 2: Day>=8
Group 1	Case	L-ARG+GHRH	AEZS-130 (macimorelin),
	Control		
Group 2	Case	AEZS-130 (macimorelin),	L-ARG+GHRH
	Control		

Prior to completion of the control subject accrual, the study was stopped. When the study was restarted, L-ARG+GHRH treatment was no longer commercially available. Subjects enrolled under amendments 4 and 5 will only receive AEZS-130 therefore the crossover aspect of the design and randomization no longer apply.

Forty-two (42) AGHD patients and 10 matched control subjects were enrolled in the original protocol. Approximately thirty (30) normal control subjects matching the AGHD patients who have completed the original protocol will be enrolled. In addition, 20 subjects; 10 adult growth hormone deficient (AGHD) patients and 10 matched normal control subjects will be enrolled. Approximately 100 subjects will be enrolled for this study.

Randomization and matching controls
In the original protocol, cases were randomized to a group (group 1 or group 2) which determined the treatment order upon entry into the study. Half of the cases received AEZS-130 first while the remaining half received L-ARG+GHRH first. Each patient's matched control was to be treated in the same order. Controls will be matched to cases based on age (± 5 yrs), sex, body mass index (± 2 kg/m2) and estrogen supplement status (if applicable).

Subjects enrolled under amendments 4 and 5 will only receive AEZS-130 therefore the crossover aspect of the design and the randomization no longer apply and there is no group or order identification required for these subjects.

Analyses population:
The **safety analysis population** will consist of all subjects who received treatment, irrespective of any post-treatment measurements for GH and IGF-1.

The **intent-to-treat analysis population (ITT)** will consist of all subjects who received a treatment and had at least one post-treatment GH measurement.

Reference ID: 3539050

The **modified intent-to-treat analysis population (mITT)** will consist of all ITT subjects that received AEZS-130 and have at least one post-treatment GH measurement between 45 and 75 minutes post-treatment. The primary efficacy analyses will be based on the MITT. For completeness, data from subjects who were treated with L-ARG+GHRH will also be summarized.

A per-protocol analysis population (PPS) is defined as all subjects without major protocol deviations. Major protocol deviations will include the following:
Cases enrolled who did not meet the entry criterion for confirmed AGHD
Not providing sufficient data to determine peak GH concentration

Subgroup definitions
Following are the subgroups of subjects that will be evaluated to assess the impact of these parameters on diagnostic assessments:

demographic (by median age, gender)
BMI (by Lean[<25], Overweight[≥25 and <30], Obese[≥30])
estrogen status

Center contributions
The cases will be enrolled from 8-10 centers with the expectation that each center will enroll ~4-5 subjects. No one center will be allowed to enroll more than a third (~17) of the total cases. Centers that enroll only 1-2 subjects may be pooled with other low enrolling centers during analysis.

The 40 additional controls and 10 additional AGHD subjects enrolled under Amendments no. 4 and 5 will be accrued from up to four centers.

Primary endpoints:

The primary endpoint for each individual is the peak GH concentration following treatment (AEZS-130). The GH concentrations are determined by assay.

The basis of the truth for the ROC analysis is the case or control status of the subject. By definition, the cases are proven AGHD subjects by meeting entry criteria. Cases without confirmed AGHD will be identified as having a major protocol deviation and excluded from the per protocol analysis set.

3.2.2 Statistical Methodologies

The statistical analysis plan (SAP) was approved on 07-Jun-2011 and the database locked on 12-Aug-2011. The statistical methods described in the SAP are summarized below and the reviewer's comments will be in italics in this Section.

Primary Efficacy Evaluation:
The primary efficacy analysis is the ROC analysis performed on the peak GH levels when the subjects were treated with AEZS-130. The case or control status of the subject will be used as

Reference ID: 3539050

"truth" for AGHD. The ROC curve is created by plotting the true positive fraction (TPF or sensitivity) vs. the false positive fraction (FPF or 1-spcificity) determined by moving the cut-point value along the peak GH level range.

The diagnostic accuracy of AEZS-130 will be established if the AUC is greater than 0.85. This will be tested with α=0.01 using the one sided hypotheses:

H0: AUC $\leq$ 0.85; HA: AUC > 0.85

The confidence interval of the AUC will be estimated using a bootstrap approach. The bootstrap process will involve drawing 1000 repeated independent samples (with replacement) from the original dataset and then estimating the AUC from the ROC analysis for each sample. The lower bound of the one-sided 99% confidence interval will be estimated directly from the 10,000 bootstrap samples using the 1%-tile of the samples (e.g., the 10th smallest sample of the 1000 samples). If the lower bound is **greater than 0.85 then** the null hypothesis will be rejected and AESZ-130 will be accepted as having diagnostic accuracy.

Secondary Efficacy Analyses include:

- As a secondary endpoint, at least 8 of the 10 newly enrolled AGHD patients should be correctly classified for a protocol pre-specified threshold of peak GH concentration level, **which is 8.5 (ng/ml)**
- The sensitivity and its 95% confidence interval will be reported for peak GH<=cut- point for all cut-points along the peak GH range for AEZS-130
- The specificity and its 95% confidence interval will be reported for peak GH<=cut-point for all cut-points along the peak GH range for AEZS-130
- The CART analysis will estimate:
    - Cut-point which minimizes the misclassification of case and control subjects
    - Optimal decision tree for AEZS-130 which incorporates age, sex and BMI
- Sensitivity and specificity will be summarized for age, gender, BMI and estrogen status subgroups containing n>10
- Logistic regression analysis with backward elimination of non-significant terms from a full model which includes subgroup covariates and covariate by AEZS-130 interactions.
- ROC analysis will be performed for the subjects that received GHRH

For the first secondary endpoint, it is not clear why the 8.5 is used here. Sensitivity should be evaluated together with specificity. The confidence intervals for the sensitivity and specificity will be explored and calculated by exact method in this review.

The sponsor did not specify the method for determining the optimal cut-points in SAP. It seems that the optimal cut-point can be selected according to the misclassification rate. CART analysis can also be used to decide the best cut-point with or without covariates. With different rules, the optimal cut-point will be different. It is not clear which rules are the proper rules for clinical use and should be included in the label. Therefore, the CART analyses will not be explored in this review.

Reference ID: 3539050

3.2.3 Patient Disposition, Demographic and Baseline Characteristics

As shown in Table 3, all enrolled subjects were included in the intent-to-treat analysis set. One AGHD patient (subject id as 0602) was enrolled, but only received L-ARG+GHRH, and discontinued without receiving macimorelin (AEZS-130), who is excluded from the mITT population.

According to the sponsor, two additional AGHD patients were excluded from the PPS analyses for the following reasons:

- One patient had only two pituitary deficiencies instead of the three required per the protocol
- Another patient did not have three hormone deficiencies or a low IGF-1 (patients were randomized in error).

Partial patients received L-ARG+GHRH because the product was withdrawn from the market during the study period.

In this review, the focus is on the pre-specified mITT population. PPS population (specified by the sponsor) will be explored in some analyses. Subjects 0609 and 0608 (cases, previously enrolled, without matched controls) are not included in the PPS population. The group of subjects in mITT population, receiving both AEZS-130 and L-ARG+GHRH, will also be explored in this review.

Table 3: Patient disposition (count) for subjects (all enrolled subjects)

Total subjects	AGHD patient N=53	Matched control N=48
Enrolled	53	48
Safety	53	48
Intent-to-treat	53	48
mITT	52	48
PPS	50	48
Received AEZS-130 in mITT	52	48
Received L-ARG+GHRH in mITT	42	10
Received AEZS-130 in PPS	50	48
Received L-ARG+GHRH in PPS	40	10

As shown in Table 4, the numbers of subjects with different sex, ethnicity, BMI status are similar in the patients and controls. The control group has slightly older patients (60% of subjects >50yr in control group and 50% subjects >50yr in patient group).

There is an imbalance in Race (4% Blacks in the patient group and 38% Blacks in the control group) in Table 4. Therefore, race may be a confounded factor for evaluating the diagnostic efficacy.

Reference ID: 3539050

Site may also be a confounded factor for evaluating the diagnostic efficacy (Table 5). For example, no AGHD patients were enrolled from Sites 11 and 17, and most of the matched controls were enrolled from Sites 09, 10, 11 and 17. Some sites (such as Site 09 and 17) enrolled more Blacks than other sites.

Table 4: Demographics and Baseline characteristics for ITT/mITT population

	AGHD patient N=52	Matched control N=48
Sex (N, %)		
Male	31 (59.6)	30 (62.5)
Female	21 (40.4)	18 (37.5)
Race (N, %)		
White	48 (92.3)	29 (60.4)
Black or African American	2 (3.9)	18 (37.5)
Asian	0	0
Other	2 (3.8)	1 (2.1)
Ethnicity (N, %)		
Hispanic or Latino	10 (19.2)	9 (18.8)
Not Hispanic or Latino	42 (80.8)	39 (81.3)
Age (N, %)		
<50 yr	26 (50)	19 (39.6)
>=50 yr	26 (50)	29 (60.4)
Mean(sd)	53 (13)	54 (13)
Median	50	54
Min-max	25-80	24-77
Estrogen status (N, %)		
None	11	12
Oral	9	6
patch	1	0
missing	31	30
BMI (kg/m2)		
Lean (<25)	7 (13.5)	7 (14.6)
Overweight (25-30)	15 (28.9)	13 (27.1)
Obese (>=30)	30 (57.7)	28 (58.3)
Mean (std)	32.1 (7.48)	31.8 (7.15)
Median	31.3	31.7
Min-Max	19.1-55.7	19.5-52.4

Reference ID: 3539050

Table 5: Distribution of subjects in mITT population by sites

	AGHD patient N=52	Matched control N=48
site (N, %)		
01	10 (19)	0
02	1 (2)	0
03	10 (19)	0
04	1 (2)	0
06	8 (15)	4 (8)
07	3 (6)	0
08	4 (8)	0
09	2 (4)	19 (40)
10	13 (25)	5 (10)
11	0	6 (13)
17	0	14 (29)

3.2.4 Results and Conclusions

Summary of the main results provided by the sponsor is shown below.

Analyses were conducted in two populations: the modified intent to treat (mITT) population and the per-protocol set (PPS). The mITT population was meant to include in the primary efficacy analysis only subjects with at least one post-treatment GH measurement between 45 and 75 minutes post-treatment. However, the mITT population included two subjects who did not meet the criteria for confirmed AGHD. Because this is an important aspect when evaluating a diagnostic agent, these two subjects were excluded from the per-protocol set (PPS) population.

The Study provided a 99% Lower Bound for the one-side CI for the AUC of .835(<0.85) for mITT population, thus the pre-specified primary analyses failed. However, for PPS population, the 99% Lower bound for the one-sided CI for the AUC is 0.8502 (>0.85).

A cut-point for the response variable (peak concentration following AEZS-130) of 2.7 ng / mL (optimal cut-point without other covariates) produced a Sensitivity of 0.81 and a Specificity of 0.92 for mITT population, a Sensitivity of 0.82 and a Specificity of 0.92 for PPS population.

The addition of a covariate (BMI) that separates the population into two groups – subjects with BMI < 30kg / M^2 versus subjects with BMI >= 30 kg / M^2 , followed by the ROC analysis performed independently on each of these groups, yielded the following results:
For mITT population, BMI-adjusted ROC AUC = 0.918. Lower bound of 99% CI is 0.845 by bootstrap method.
For PPS population, BMI-adjusted ROC AUC = 0.925. Lower bound of 99% CI is 0.853 by bootstrap method.

Reference ID: 3539050

Subgroup analyses were conducted for the PPS population.
For the BMI < 30kg / M^2 group:
AUC = .913, with "Optimal" cut-point 6.8 ng / mL
For the BMI >= 30kg / M^2 group:
AUC = .937, with "Optimal" cut-point 2.7 ng / mL

Overall, Sensitivity=0.86, specificity=0.92 (using 6.8 ng/mL in BMI<30kg / M^2 group and 2.7 ng/mL in the BMI>=30kg / M^2 group). Thus, the specification of a good cut-point in the peak concentration following AEZS-130 can depend on the subject's BMI.

The reviewer's results are consistent with the sponsor's results above. **The reviewer's findings** are presented below.

Summary of the peak growth hormone concentration (peakcon)

Summary of the peak growth hormone concentration (ng/mL) following Macimorelin (AEZS-130) administration in the mITT population is shown in Table 6. The mean, median and range of the peak concentration are very different in the patient group and the control group. The distributions of the peakcon in the two groups are very different.

Table 6: Summary of growth hormone concentration in AGHD patient and matched controls

Analysis population	AGHD patient				Matched controls			
	N	Mean (sd)	Median (Q1-Q3)	Min-max	N	Mean (sd)	Median (Q1-Q3)	Min-max
mITT	52	2.70 (6.22)	0.55 (0.095-1.5)	0.025-33.0	48	17.71 (19.11)	10.5 (5.6-22.6)	0.85-94.0

ROC analyses

ROC AUC results are presented in Table 7. **Bold value is from the pre-specified primary analysis. The lower bound of 99% 1-sided confidence interval is 0.83, which is less than the pre-specified threshold 0.85. The sponsor failed on the pre-specified primary analysis.**

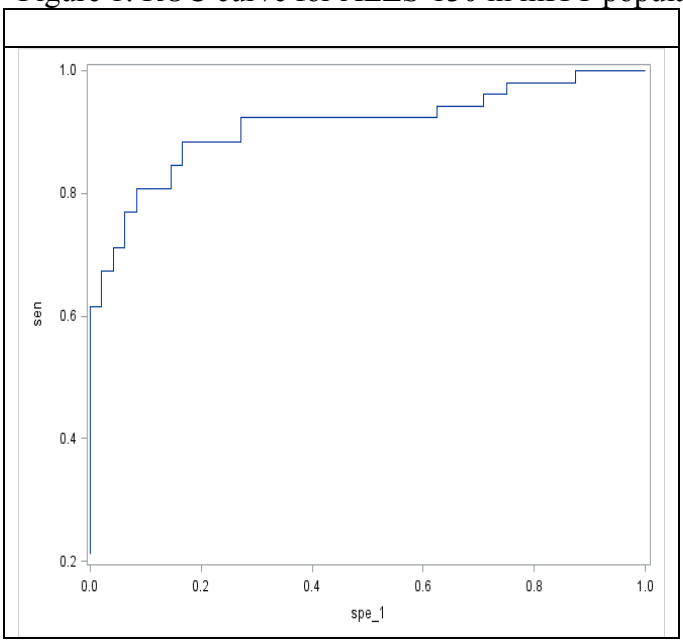
The PPS analyses had slightly higher values of lower bound of 1-sided CIs compared with the mITT analyses. The ROC curve for AEZS-130 is shown in Figure 1.

Reference ID: 3539050

Table 7: ROC analyses for mITT population

	mITT population	PPS
ROC AUC	0.912	0.923
Lower bound of 1-sided		
99% CI	**0.830**	0.851
97.5% CI	0.845	0.864
95% CI	0.858	0.875

Figure 1: ROC curve for AEZS-130 in mITT population (AUC=0.913)



Secondary analyses: Sensitivity and specificity evaluation by cut-points

The sensitivity and specificity values, and percent of misclassification (disagreement) for different cut-points are also presented in Table 8 (only percent of misclassification<.20% were presented). The cut-points are the possible values of the peak hormone concentration in ng/mL.

The possible cut-points could be values between 1 and 7 ng/mL. If higher specificity is preferred by clinical practice, smaller cut-points could be selected.

For mITT population, sensitivity is 0.81 (95% CI: 0.68, 0.90) and specificity is 0.92 (95% CI: 0.80, 0.98) with cut-point as 2.7 ng/mL (sponsor selected this cut-point as the optimal cut-point without other covariates). 2.7 is one of the cut-points with the smallest percent of disagreement value (14%) shown in Table 8. If the peak concentration is <=2.7 ng/mL then the subject is abnormal, otherwise, the subject is normal.

Reference ID: 3539050

Table 8: Sensitivity, specificity, and percent of disagreement (in %) for mITT population with percent of disagreement <20%

Obs	sensitivity	specificity	percent of disagreement	Cut-point in ng / mL
1	65.4	97.9	19	0.92
2	67.3	97.9	18	1.00
3	71.2	95.8	17	1.30
4	75.0	93.8	16	1.40
5	76.9	93.8	15	1.60
6	78.8	91.7	15	2.10
7	80.8 (67.5, 90.4)	91.7 (80, 97.7)	**14**	**2.70**
8	80.8	87.5	16	3.00
9	80.8	85.4	17	3.10
10	82.7	85.4	16	3.30
11	84.6	85.4	15	3.50
12	84.6	83.3	16	3.80
13	86.5	83.3	15	4.00
14	88.5 (76.6, 95.6)	83.3 (69.8, 92.5)	**14**	**4.30**
15	88.5	79.2	16	4.50
16	88.5	77.1	17	5.20
17	88.5	75.0	18	5.50
18	88.5	72.9	19	5.70
19	90.4	72.9	18	6.80
20	92.3	72.9	17	7.00
21	92.3	70.8	18	7.30

Note: values in () are 95% CIs for some sensitivity and specificity values

Secondary analysis: sensitivity in newly-enrolled and previously enrolled AGHD patients

As shown in Table 9, for both sensitivity and specificity using the pre-specified cut-point 8.5 ng/mL, the performance for the new enrolled patients is worse than the previously enrolled subjects (80%<95% for sensitivity and 53%<90% for specificity). In addition, the specificity (30%) is worse if we only include the newly enrolled matched controls in the analyses.

Reference ID: 3539050

The performance for the newly enrolled and previously enrolled subjects is not consistent. In addition to the pre-specified cut-point 8.5, another cut-point 2.7 was also used and the performance is shown in Table 9. The similar pattern is observed.

The specificity is very low (point estimates about 50%) using the 8.5 cut-point for the newly enrolled subjects. It is also not clear if the relatively better performance with cut-point 2.7 can be shown in another study.

Table 9: sensitivity and specificity for newly enrolled subjects and all subjects

Cut-point	Groups	Sensitivity	CI in %	Specificity	CI in %
8.5 ng / mL	Newly enrolled	**8/10=80%**	**(44,98)**	**3/10=30%**	**(7, 65)**
		8/10=80%	(44,98)	20/38=53%	(36, 69)
	Previously enrolled	40/42=95%	(83, 99)	9/10=90%	(55, 100)
	All	48/52=92%	(82,98)	29/48=60%	(45, 74)
2.7 ng / mL	Newly enrolled	**7/10=70%**	**(35, 93)**	**8/10=80%**	**(44, 97)**
		7/10=70%	(35, 93)	34/38=89%	(75, 97)
	Previously enrolled	35/42=83%	(68, 93)	10/10=100%	(69, 100)
	All	42/52=81%	(68, 91)	44/48=92%	(80, 98)

Results with Bold include 10 newly enrolled matched controls for the specificity evaluation.

Secondary analyses: ROC analyses with adjustment (mITT population)

In the logistic model with peak concentration (peakcon) following Macimorelin (AEZS-130) administration, age, baseline BMI (baseBMI), sex, race, only peakcon is highly significant with nominal p-value <.0001 in most of the models.

Models with site cannot be fitted because of small sample size. Subjects are also grouped into two BMI groups. If baseBMI<30 then BMIgrp=0, otherwise BMIgrp=1;

ROC AUC values with and without adjustment of other covariates have similar ROC AUC values around 0.9. Without adjustment, ROC AUC is 0.913 (point estimate). With adjustment for BMI group (BMI<30 or >=30) and the interaction between Peakcon and BMI group, ROC AUC is 0.918; with adjustment for RACE, ROC AUC is 0.925.

Secondary analyses: ROC analyses for the patients receiving the comparator drug L-ARG+GHRH

The analyses are conducted for subjects (52 subjects) with both Macimorelin (AEZS-130) and L-ARG (the comparator). The AUC values (point estimates) for both drugs are above 0.9. The performance of AEZS-130 is slightly better than the comparator (AUC 0.986 vs. 0.929). Note that there are only 10 controls in the data, the sample size is small. The 95% CI confidence intervals were obtained by Mann-Whitney method.

Table 10: AUC for model with different drugs (test and comparator)

Model	AUC (95%CI) mITT population
Peak concentration only (LARG+GHRH)	0.929 (0.848, 1)
Peak concentration only (Macimorelin (AEZS-130))	0.986 (0.961, 1)

Secondary analyses: sensitivity and specificity analyses for the patients receiving the comparator drug L-ARG+GHRH

In addition to the ROC analyses, the sensitivity and specificity of the two drugs with some cut-points were calculated and are shown in Table 11. The optimal points were selected as the ones with the smallest percent of disagreement.

The optimal cut-points (with less percent of disagreement and higher specificity if the percent of disagreement is the same) for Macimorelin (AEZS-130) and L-ARG (arginine)+GHRH are different. The results on sensitivity and specificity using the optimal cut-points are similar for Macimorelin (AEZS-130) and LARG+GHRH, and Macimorelin (AEZS-130) has slightly higher sensitivity and specificity.

Table 11: sensitivity and specificity, and percent of disagreement for subjects with the comparator

Cut-point on the peak concentration	Drug	sensitivity	CI in %	Specificity	CI in %	Percent of disagreement
2.7 ng / mL	LARG+GHRH	29/42=69%	(53, 82)	9/10=90%	(55, 100)	27
	Macimorelin (AEZS-130)	35/42=83%	(68, 93)	10/10=100%	(69, 100)	13.5
4.3 ng / mL (optimal for AEZS-130)	Macimorelin (AEZS-130)	38/42=90.5%	(77, 97)	10/10=100	(69, 100)	7.7
7.4 ng / mL (optimal for LARG)	LARG+GHRH	36/42=85.7%	(71, 95)	9/10=90	(55, 100)	13.5

Reference ID: 3539050

3.3 Evaluation of Safety

There is no major safety issue for this product (For more details please see clinical review).

4. FINDINGS IN SPECIAL/SUBGROUP POPULATION

In this Section, subgroup analyses for mITT population as exploratory analyses were presented. Only reviewer's results are presented below.

Note that no conclusions can be drawn from the subgroup analyses due to lack of representation and limited sample size. Further exploration should be conducted for some subgroups of interest.

4.1 Gender, Race, and Age

Sensitivity and specificity were explored by gender, race and age, using the cut-point selected by the sponsor (peak concentration 2.7 ng/mL). The results are shown in Table 12. The accuracy in terms of sensitivity and specificity is slightly better for subjects with age>=50 vs. subjects with age<50 (sensitivity 85% vs. 77%, specificity 93% vs. 89%); for Blacks vs. Whites (sensitivity 100% vs. 79%, specificity 94% vs. 90%). The specificity is better in Male vs. Female (97% vs. 83%).

Table 12: Sensitivity (sen) and specificity(spe) by subgroups (in %).

		Sensitivity	Specificity
Age	<50 (n=45)	77 (n=26)	89 (n=19)
	>=50 (n=55)	85 (n=26)	93 (n=29)
Race	White n=77	79 (n=48)	90 (n=29)
	Black n=20	100 (n=2)	94 (n=18)
	Other n=3	100 (n=2)	100 (n=1)
Sex	Male n=61	81 (n=31)	97 (n=30)
	Female n=39	81 (n=21)	83 (n=18)

n is sample size.

4.2 Other Special/Subgroup Populations

From earlier exploration, BMI may be interacted with the test for evaluating the diagnostic efficacy, and there is imbalance observed in sites. Therefore, the accuracy in term of sensitivity and specificity (using the cut-point 2.7 ng/mL) were evaluated by baseline BMI values and the sites. ROC analyses were conducted by BMI, but not sites because of small sample size in many sites.

Reference ID: 3539050

Subgroup ROC analyses by BMI
The ROC AUC is 0.892 and 0.939 for the subjects with baseBMI<30kg / M^2 and >=30kg / M^2, respectively.

By Man-Whitney method, the 95% CI for the AUC is (0.791, 0.993) and (0.872, 1) for the subjects with baseBMI<30kg / M^2 and >=30kg / M^2 , respectively.

The 1-sided 99%, 97.5% and 95% CI lower bound of the ROC AUC for the subgroups by bootstrap method is 0.754, 0.78, 0.80 for subjects with baseBMI<30kg / M^2 , respectively; 0.851, 0.866, 0.880 for subjects with baseBMI>=30kg / M^2 , respectively.

The confidence intervals from the Mann-Whitney and bootstrap methods are consistent.
Using the combined criteria to determine disease or not disease, the overall sensitivity is 92%, specificity is 85%.:

- if baseBMI<30kg / M^2 and peak concentration <=**6.8 ng/mL** then flag=**positive (disease)**;
- if baseBMI<30kg / M^2 and peak concentration>**6.8 ng/mL** then flag=**negative (not disease)**;
- if baseBMI>=30kg / M^2 and peak concentration<=**2.7** ng/mL then flag=**positive (disease)**;
- if baseBMI>=30kg / M^2 and peak concentration>**2.7** ng/mL then flag=**negative (not disease)**;

The sensitivity and specificity values for the subgroups are shown in Table 13. The performance is slightly better for subjects with BMI>=30 vs. BMI<30 (sensitivity 83% vs. 77%, specificity 96% vs. 85%).

Specificity can only be evaluated in 4 sites (100%, 95%, 100%, and 79% for sites 06, 09, 10, and 17). The sample size for sites 06 and 10 are small (n<=5). For site 10, the sensitivity value is low (69%) even though the specificity is 100%. The performance varies in different sites.

Reference ID: 3539050

Table 13: Sensitivity (sen) and specificity (spe) point estimates by BMI and site (in %) using 2.7 ng/mL as the cut-point

		Sensitivity	Specificity
BMI	<30 n=42	77 (n=22)	85 (n=20)
	>=30 n=58	83 (n=30)	96 (n=28)
site	01 n=10	80 (n=10)	No data
	02 n=1	100 (n=1)	No data
	03 n=10	80 (n=10)	No data
	04 n=1	100 (n=1)	No data
	06 n=12	87.5 (n=8)	100 (n=4)
	07 n=3	100 n=3	No data
	08 n=4	75 n=4	No data
	09 n=21	**100 n=2**	**95 n=19**
	10 n=18	**69 n=13**	**100 n=5**
	11 n=6	100 n=6	No data
	17 n=14	No data	**79 n=14**

n is sample size.

5. SUMMARY AND CONCLUSIONS

5.1 Statistical Issues

1) Statistical significance was not achieved in the pre-planned primary analysis of the study. For the pre-planned primary analysis, the lower bound of the 99% confidence interval for the ROC AUC is <0.85, the pre-specified threshold.

2) There is a significant imbalance in Race between the AGHD group and the matched control group. Two of 52 AGHD patients (4%) and 18 out of 48 matched controls (38%) are Black or African American. Race may be a confounded factor for evaluating the diagnostic efficacy.

3) There is a significant imbalance in study site between the AGHD group and the matched control group. For example, no AGHD patients were enrolled from Sites 11 and 17, and most of the matched controls were enrolled from Sites 09, 10, 11 and 17. Site may be a confounded factor for evaluating the diagnostic efficacy.

4) The determination of the optimal cut-point is a challenge. Different approaches may lead to different cut-points. In addition, the selected cut-point was based on the same data that was used to determine its sensitivity and specificity. Therefore, the estimates of sensitivity and specificity may tend to overstate the truth. It is not clear if the sensitivity and specificity obtained in Study AEZS-130-047 can be reproduced in another study.

Reference ID: 3539050

5) For both sensitivity and specificity using cut-point 8.5 (pre-specified by the sponsor in the protocol) and 2.7 (optimal cut-point determined from the data by the sponsor), the performance for the newly enrolled patients is worse than the previously enrolled subjects. The results are not consistent (previously vs. newly enrolled).

6) It is not clear if the matched control group includes subjects belonging to the intended patient population, who would seek for diagnosis of possible AGHD and would benefit from the test. The performance (ROC AUC, sensitivity and specificity) of Macimorelin for the intended patient population is not known.

5.2 Collective Evidence

This NDA submission includes five studies. Three studies enrolled only healthy volunteers and one study is ongoing. Therefore, only one study (Study AEZS-130-047) is relevant and all the evidence (summarized below) is from Study AEZS-130-047.

The Study provided a 99% Lower Bound for the one-side CI for the AUC of .83 (less than the pre-specified threshold of 0.85) for mITT population, thus the pre-specified primary analyses failed. For PPS population defined by the sponsor after obtaining the data, the 99% Lower bound for the one-sided CI for the AUC is 0.851 (>0.85).

With the cut-points from 2 ng/mL to 7 ng/.mL, the range of sensitivity and specificity values are from 80% to 90%, and percent of misclassification (disagreement between the test and the truth) ranges from 14% to 20% for the mITT population. A cut-point for the response variable (peak concentration following AEZS-130) of 2.7 ng / mL produced a Sensitivity of 0.81 (95 CI: 0.68, 0.90) and a Specificity of 0.92 (95% CI: 0.80, 0.98) for mITT population.

52 subjects (42 AGHD patients + 10 matched controls) were enrolled previously under the original protocol and received both Macimorelin (AEZS-130) and L-ARG+GHRH (the comparator). 48 additional subjects (10 AGHD patients and 38 matched controls) were enrolled under the SPA agreed by the sponsor and the Agency and received Marcimorelin only because the comparator is not available during the study period.

The performance (see details in Table 9) for the new enrolled patients is worse than the previously enrolled subjects (80%<95% for sensitivity and 53%<90% for specificity using the pre-specified cut-point 8.5 ng/mL). In addition, the specificity (30%) becomes worse if we only include the 10 newly enrolled matched controls for the 10 newly enrolled patients in the specificity analysis. The similar pattern is observed for another cut-point 2.7ng/mL determined using the data. The results are not consistent for subjects enrolled from different study portions.

The addition of a covariate (BMI) that separates the population into two groups – subjects with BMI < 30kg / M^2 versus subjects with BMI >= 30 kg / M^2 , followed by the ROC analysis performed independently on each of these groups, yielded the following results:

For mITT population, BMI-adjusted ROC AUC = 0.918. Lower bound of 99% CI is 0.845 by bootstrap method. Using the combined criteria of BMI with cut-point **30kg / M^2** and the peak concentration cut-point 6.8 ng/mL in subjects with BMI<30 and 2.7 ng/mL for subjects with BMI>=30, the sensitivity is 0.86 and specificity is 0.92 for the mITT population.

Subgroup analyses were conducted for the mITT population.
For the BMI < 30kg / M^2 group: AUC = .892, lower bound of 99% CI is 0.975. Sensitivity is 77% and specificity is 85% with cut-point 6,8 ng/mL
For the BMI >= 30kg / M^2 group: AUC = .939, lower bound of 99% CI is 0.851. Sensitivity is 83% and specificity is 96% with cut-point 2.7 ng/mL.

For the 52 subject with both Macimorelin (AEZS-130) and L-ARG+GHRH (the comparator). The AUC value (95% CI) is 0.986 (0.961, 1) and 0.929 (0.848, 1) for Macimorelin and L-ARG, respectively. With the optimal cut-point 4.3 ng/mL for Macimorelin, the sensitivity is 0.91 (95% CI: 0.77, 0.97), and the specificity is 1.00 (95% CI: 0.69, 1.00). With the optimal cut-point 7.4 ng/mL for L-ARG, the sensitivity is 0.86 (95% CI: 0.71, 0.95), and the specificity is 0.90 (95% CI: 0.55, 1.00).

The intended patient population is not clear and the performance of Macimorelin for the intended patient population is unknown from the pivotal study.

5.3 Conclusions and Recommendations

The statistical results in terms of accuracy do not provide enough evidence to support the claims proposed in this NDA. The sponsor failed in the pre-specified primary ROC AUC analysis. The ROC AUC results may be biased because of possible confounding of the test with race and study site. The sensitivity and specificity results are not consistent in different subject groups (previously vs. newly enrolled). Different cut-points can be selected according to the rules (not pre-specified) and the cut-points are selected based on the single data with 100 subjects, which may not be reproducible in another study. We recommend CR with additional confirmatory study with pre-specified cut-point and intended patient population.

5.4 Labeling Recommendations

We recommend CR for this product. However, if this product is approved, we recommend the following changes in Indication and Section 14 (Clinical studies) in the label.

(b) (4)

This is a representation of an electronic record that was signed electronically and this page is the manifestation of the electronic signature.

/s/

LAN HUANG
07/08/2014

MARK D ROTHMANN
07/09/2014
I concur

THOMAS J PERMUTT
07/09/2014
I concur.

STATISTICS FILING CHECKLIST FOR A NEW NDA/BLA

NDA Number: 205598/0000 **Applicant: Aeterna Zentaris GmbH** **Stamp Date: 11-05-2013**

Drug Name: Macimorelin Acetate **NDA/BLA Type: Original NDA** **Submission Date: 11-04-2013**

On **initial** overview of the NDA/BLA application for filing:

	Content Parameter	Yes	No	NA	Comments
1	Index is sufficient to locate necessary reports, tables, data, etc.	x			
2	ISS, ISE, and complete study reports are available (including original protocols, subsequent amendments, etc.)			**x**	One study
3	Safety and efficacy were investigated for gender, racial, and geriatric subgroups investigated (if applicable).		x		Matching was performed on three criteria – age, sex/estrogen supplement and BMI
4	Data sets in EDR are accessible and do they conform to applicable guidances (e.g., existence of define.pdf file for data sets).				Could not open Define.xml

IS THE STATISTICAL SECTION OF THE APPLICATION FILEABLE? ____Yes____

If the NDA/BLA is not fileable from the statistical perspective, state the reasons and provide comments to be sent to the Applicant.

(1) Please provide subgroup analysis by sex, age and race (and geographic region if applicable) or provide a justification as to why these analyses are not included.

(2) We cannot open define.xml file. Please provide a corresponding pdf file.

Studies Submitted

AEZS-130-047

Studies to be Reviewed

AEZS-130-047

Please identify and list any potential review issues to be forwarded to the Applicant for the 74-day letter.

Content Parameter (possible review concerns for 74-day letter)	Yes	No	NA	Comment

File name: 5_Statistics Filing Checklist for a New NDA_BLA110207

STATISTICS FILING CHECKLIST FOR A NEW NDA/BLA

Designs utilized are appropriate for the indications requested.				
Endpoints and methods of analysis are specified in the protocols/statistical analysis plans.				
Interim analyses (if present) were pre-specified in the protocol and appropriate adjustments in significance level made. DSMB meeting minutes and data are available.			x	
Appropriate references for novel statistical methodology (if present) are included.		x		
Safety data organized to permit analyses across clinical trials in the NDA/BLA.				
Investigation of effect of dropouts on statistical analyses as described by applicant appears adequate.			x	One patient missed the primary effic. Value after treatment

Further Information Request in 74-day letter

Japobrata Choudhury, Ph.D. 12-31-13

Reviewing Statistician Date

Mark Rothmann, Ph.D.

Supervisor/Team Leader Date

File name: 5_Statistics Filing Checklist for a New NDA_BLA110207

This is a representation of an electronic record that was signed electronically and this page is the manifestation of the electronic signature.

/s/

JAPOBRATA CHOUDHURY
12/31/2013

MARK D ROTHMANN
12/31/2013
concur