scans. Each blinded reader evaluated the scans independently of the other two readers. The rating specified by two or more readers (majority) of each scan study was the rating used for analysis.

Each blinded reader was provided with the two scan series for a patient. One scan series included a Thyrogen 48 hour scan(s) and a Withdrawal 48 hour scan(s). When available, the 72 hour scan was also evaluated. Each scan series was reviewed separately within a patient. The evaluation of the 48 hour scan was completed before the reader moved on to the evaluation of the combination of the 48 and 72 hour scans from the same scan series, i.e. the 48 and 72 hour Thyrogen scans and the 48 and 72 hour Withdrawal scans. Upon completion of the evaluation of scans from the one scan series, the reader evaluated the other scan series.

The reader evaluated the scan using a cancer classification system (see Table 6.2) and scored the scan according to the highest class of cancer observed. The number, location on an anatomical diagram, and distribution of lesions observed within each classification was recorded. An assessment for evidence of possible artifact (e.g. saliva in the esophagus, urine or saliva contamination of the skin) was made. This evaluation included any additional views which were taken at the time of the scans to investigate the possibility of an artifactual focus of uptake. Any focus of uptake determined to be an artifact was documented. If the reader was unable to determine the class observed on the scan series because he could not rule out artifactual uptake, it was documented.

For patients with either post-ablation or post-therapy scans, the reader determined if this scan confirmed the cancer classification seen on the 48 hour scan and the combination of the 48 and 72 hour scans from both scan series. If the classification of a scan was not confirmed, the reader indicated why.

For patients with uptake limited to the thyroid bed, the reader evaluated the most recent pre-study scan, when available, to attempt to delineate whether the uptake in this area was thyroid remnant or local cancer recurrence. When such a determination was possible, the type of tissue was specified.

Study scans were designated to be evaluated as a consensus panel, consisting of the same three independent readers, under the following conditions:

- The readers did not reach a consensus on the cancer classification rating of a study scan.
- A majority of the readers recommended that a scan be sent to panel because of the difficulty in determining a rating for the scan.
- A majority of the readers responded that a within classification difference in the number and distribution of lesions between the within patient scan pair could potentially change or alter the clinical management of the patient.

Additional unmatched 48 or 72 hour views might have been used to help the readers reach a consensus. If this panel evaluation failed to achieve a consensus, it was documented and the data for this patient was not included in the efficacy analysis.

5.3 Key Differences Between Studies

The differences between Study TSH95-0101 and Study TSH92-0601 are presented in Volume 1.58 on pages 17 to 20. The key differences summarized below are the features found in Study TSH95-

0101 that are not found in Study TSH92-0601.

- The exclusion of hemi or partial-thyroidectomy patients.
- The ^{131}I dose was increased to 4 mCi for all patients instead of a range of 2 to 4 mCi;
- A 72 hour scan might have been done to help delineate an actual focus of uptake in thyroid remnant or malignant cancer.
- The definition of equivalent scan was changed from both scans in the scan pair showing the same stage, number, and distribution of lesions to both scans showing the same distribution, as described in the "classification" scheme. A superior WBS showed the presence of thyroid remnant tissue and/or the more extensive cancer that was not seen on the other WBS. The location of the relevant focus(I) of uptake was identified in both the Thyrogen and Withdrawal scan series, and the number of lesions observed was recorded.
- No separate assessment of each individual WBS prior to the within-patient comparison of WBS was done. The sponsor stated that the independent reviewers in Study TSH92-0601 felt that the separate assessment of each individual WBS prior to the within-patient comparison of WBS introduced a level of unnecessary complexity. Therefore, only a within-patient, paired, comparison of the Thyrogen and Withdrawal scans was done (Volume 1.58, page 19).
- Post-ablation or post-therapy radioiodine scans or histology reports from surgery were required. Scan results and histology reports were used to confirm the presence of cancer.
- The results of the patient's most recent WBS obtained in routine follow-up prior to entering the study was used, if available, to document previously noted remnant tissue, residual, or recurrent cancer.

6.0 STUDY VARIABLES

6.1 Study TSH92-0601

6.1.1 Demographics and Safety Variables

The gender, age, height, and weight of the patient were recorded. Safety data included vital signs (temperature, blood pressure and pulse rate), medical history, THST regimen, concomitant medication, pregnancy status for females, hematology, blood chemistry, thyroid function tests (serum TSH, Tg, and Tg antibody level measurement), and patient immune response to Thyrogen.

6.1.2 Primary Efficacy Endpoints

The primary efficacy endpoint was the concordance or discordance between the 48 hour Thyrogen and 48 hour Withdrawal scans for the cancer classification as evaluated by the blinded readers. Table 6.1 presents the cancer classification system used by the blinded readers. (The principal investigator also evaluated the scans but this data was not used in the efficacy analyses.)

There were two agreement measures used in the efficacy analyses. The first was the separate scan evaluation within patient. The rating for a Thyrogen scan for one patient was based on the majority of at least two of the three blinded reader ratings of the scan. If the patient's scan was sent for panel review, the panel rating was used.

Using the data from this procedure, the Thyrogen and Withdrawal scan scores were compared and if both were found to document an identical class, the two scans were considered equivalent, provided a clinically important within classification discordance between the scans had not been documented. The WBS from a scan series which documented the presence of a higher (more severe) classification of cancer was considered superior to the WBS from the other series. The three possible outcomes for each patient were as follow:

- The Thyrogen scan was rated higher than Withdrawal scan.
- The Thyrogen scan was equivalent to the Withdrawal scan.
- The Thyrogen scan was rated lower than Withdrawal scan.

The second measure was based on the side-by-side comparison of the Thyrogen and Withdrawal scans by each blinded reader. The reader was asked to determine which of the scans was superior to the other or if they were equivalent. Then a majority of two or more readers of this data was the final outcome used in the efficacy analysis.

Table 6.1
Study TSH92-0601:
Scan Cancer Classification System

Stage	Description
0	No evidence of thyroidal uptake
1	Remnants and/or thyroid cancer located within the thyroid bed. 1A: Thyroid remnant tissue 1B: Thyroglossal duct / pyramidal lobe remnant
2	Uptake outside of the thyroid bed and limited to the neck region. Therefore, local and/or nodal metastases.
3	Distant metastases: mediastinum and/or lungs with or without neck sites. 3A: Diffuse pulmonary disease 3B: Nodular pulmonary disease 3C: Nodular disease in the mediastinum
4	Distant metastases: outside of the chest and neck regions 4A: Bone metastases 4B: Brain metastases 4C: Hepatic disease 4D: Other

Source: Figure 2a, Volume 1.52, page 43.

6.1.3 Secondary Efficacy Endpoints

There were three secondary efficacy endpoints. The first was the percent ^{131}I uptake in all foci indicative of remnant and thyroid cancer tissue measured at the time of each scan. The second was the blinded reader evaluation of the presence or absence of normal sites of ^{131}I concentration (i.e., in the sinus, salivary glands, GI tract, and bladder) on Thyrogen and Withdrawal scans. The third endpoint was based on two hypothyroid symptoms assessment instruments. These instruments were the Billewicz Scale, a physician rated scale of the signs and symptoms of hypothyroidism, and the short form Profile of Mood States (POMS) Scale, a patient self-administered scale to assess areas of tension-anxiety, depression-dejection, anger-hostility, confusion-bewilderment, vigor-activity, and fatigue-inertia. The hypothyroid assessment instruments were administered at baseline and at the 48 hour Thyrogen and 48 hour Withdrawal scan visits prior to scan.

6.2 Study TSH95-0101

6.2.1 Demographics and Safety Variables

The gender, age, height, and weight of the patient were recorded. Safety data included vital signs (temperature, blood pressure and pulse rate), medical history, THST regimen, concomitant medication, pregnancy status for females, hematology, blood chemistry, thyroid function tests (serum TSH, Tg, and Tg antibody level measurement), and patient immune response to Thyrogen.

6.2.2 Primary Efficacy Endpoints

There were two primary efficacy endpoints. The first was the concordance or discordance between the 48 hour Thyrogen and 48 hour Withdrawal scans for the cancer classification as evaluated by the blinded readers. Table 6.2 presents the cancer classification system used by the blinded readers. This was done for each of the two arms of the study. (The principal investigator also evaluated the scans but this data was not used in the efficacy analyses.)

The WBS from a scan series which documented the presence of a higher (more severe) classification of cancer was considered superior to the WBS from the other series. The three possible outcomes for each patient were as follow:

- The Thyrogen scan was rated higher than Withdrawal scan.
- The Thyrogen scan was equivalent to the Withdrawal scan.
- The Thyrogen scan was rated lower than Withdrawal scan.

The second primary efficacy endpoint was based on hypothyroid symptoms assessed by the physician using the Billewicz scale. The Billewicz Scale is a physician rated scale used to document the presence or absence of the signs and symptoms of hypothyroidism. The instrument was administered at baseline and at the 48 hour Thyrogen and 48 hour Withdrawal scan visits prior to scan.

Table 6.2
Study TSH95-0101:
Scan Cancer Classification System

Description	Class	Criteria
No Uptake	0	No evidence of post-thyroidectomy thyroid remnant, well-differentiated thyroid cancer within the thyroid bed, or metastases
Uptake Limited to the Thyroid Bed	1	Evidence of well-differentiated thyroid cancer or persistent remnant limited to the thyroid bed
Uptake Outside of Thyroid Bed but Limited to the Neck Region (Exclusive of Class 1)	2	Evidence of well-differentiated thyroid cancer local metastases
Uptake evident	2A	Solitary focus uptake
Uptake evident	2B	Multiple focus uptake
Uptake in the Chest	3	Evidence of distant metastases
	3A	Uptake in mediastinum but not in the lungs
	3B	Nodular foci of uptake in the lungs
	3C	Diffuse uptake in the lungs
	3D	Any combination of 3A, 3B, and/or 3C
Uptake Outside of the Neck and Chest Areas	4	Evidence of distant metastases
	4A	Solitary focus in the skeleton
	4B	More than one focus in the skeleton
	4C	One or more foci of uptake in the liver tissue
	4D	One or more foci of uptake in the brain tissue
	4E	Any combination of 4A or 4B with 4C or 4D

Source: Figure 3H, Volume 1.56, page 42.

6.2.3 Secondary Efficacy Endpoints

Secondary efficacy endpoints were assessed for each of the two arms of the study.

Several sets of diagnostic utility endpoints were calculated for several test modalities with respect to two reference standards. Each set consisted of prevalence, sensitivity, specificity, positive and negative predictive values, and accuracy. Table 6.3 presents the test modalities and reference standards used.

Table 6.3
Per Protocol Diagnostic Utility Analysis (For the Detection of Metastatic Cancer Only)

Diagnostic Test Evaluated	Reference Standard Used	Reference Standard 1	Reference Standard 2
THST Tg level	1,2	Withdrawal post therapy scan class ≥ 2	Withdrawal post therapy scan class ≥ 2 OR a negative post therapy scan with Withdrawal Tg ≥ 10 ng/mL OR if no post therapy scan, a Withdrawal Tg ≥ 10 ng/mL and a decision by the physician to treat the patient (e.g. ^{131}I therapy, surgical dissection, external radiation, etc.).
Thyrogen Tg level	1,2		
Withdrawal Tg level	1		
Thyrogen Diagnostic WBS	2		
Withdrawal Diagnostic WBS	2		
Combination of a Thyrogen Diagnostic WBS and Tg level	1,2		
Combination of a Withdrawal Diagnostic WBS and Tg level	1		

Source: Figure 4 from Section 5.3: Statistical Methods and Analysis Plans, Volume 1.59, page 286.

Another secondary endpoint was based on an assessment of the patient's quality of life (QOL) using the SF-36 QOL, a validated testing instrument. The SF-36 QOL instrument is a patient self-administered scale is a generic quality of life instrument which measures eight health concepts, i.e. physical functioning, role limitations due to physical health problems, bodily pain, general health, vitality, social functioning, role limitations due to emotional problems, and mental health. The instrument was administered at baseline and at the 48 hour Thyrogen and 48 hour Withdrawal scan visits prior to scan.

6.2.4 Tertiary Efficacy Endpoints

The following were tertiary endpoints: TSH level after Thyrogen administration and during Withdrawal; the mean percent ^{131}I thyroidal uptake; the kinetic profile of the Tg response after Thyrogen administration; and additional analyses.

No further mention of these tertiary endpoints will be made in this review. It will focus on the primary and secondary endpoints.

7.0 STUDY HYPOTHESES AND SAMPLE SIZE

7.1 Study TSH92-0601

7.1.1 Study Hypotheses

7.1.1.1 Primary Endpoint

Comparison of Scan Cancer Classification

The protocol did not explicitly state a statistical hypothesis but analysis was based on the proportions:

- The Thyrogen WBS classification was higher than or equal to the Withdrawal WBS

classification.

- The Thyrogen WBS classification was lower than the Withdrawal WBS classification.

7.1.2 Sample Size

The protocol stated that the sample size of a minimum of 100 evaluable patients was selected so that 90% of the pairs were clinically equivalent and the disagreeing pairs were evenly divided as to which scan was higher, then the confidence interval would imply that the Thyrogen scan was at least clinically equivalent to (or higher than) the Withdrawal scan in 90% of patients.

To account for patients who did not complete study or for inadequate scans, up to 150 patients were to be enrolled.

7.2 Study TSH95-0101

APPEARS THIS WAY
ON ORIGINAL

7.2.1 Study Hypotheses

7.2.1.1 Primary Endpoints

Comparison of Scan Cancer Classification Within Treatment Arm

The protocol did not explicitly state the hypothesis but it was based on the proportions for the following two groups in each treatment arm:

- The Thyrogen WBS classification was higher than or equal to the Withdrawal WBS classification.
- The Thyrogen WBS classification was lower than the Withdrawal WBS classification.

Comparison of Scan Cancer Classification between Treatment Arms

The hypotheses to determine whether Arm I results were significantly different from Arm II results were as follow:

H_0: $\theta_{\text{Arm II}} = \theta_{\text{Arm I}}$; versus

H_1: $\theta_{\text{Arm II}} \neq \theta_{\text{Arm I}}$

where θ was the proportion of discordant scans where the Thyrogen WBS was classified higher than the Withdrawal WBS for each arm.

Hypothyroid Symptoms

Using data from the Billewicz scale, the hypotheses to be tested were:

H_0: There is no difference in hypothyroid symptoms after Thyrogen administration and during THST Withdrawal

H_1: Hypothyroid symptoms are different between after Thyrogen administration and during THST Withdrawal.

7.2.1.2 Secondary Endpoints

Diagnostic Utility of Thyrogen Tg Testing Compared to THST (Baseline) Tg Testing

The hypotheses to be tested were:

H_0: Thyrogen Tg testing correctly diagnoses the presence or absence of cancer no more

often than THST Tg testing.

H_1: Thyrogen Tg testing correctly diagnoses the presence or absence of cancer more often than THST Tg testing.

Diagnostic Utility of Thyrogen Tg Testing Compared to Withdrawal Tg Testing

The hypotheses to be tested were:

H_0: Thyrogen Tg testing correctly diagnoses the presence or absence of cancer no more often than Withdrawal Tg testing.

H_1: Thyrogen Tg testing correctly diagnoses the presence or absence of cancer more often than Withdrawal Tg testing.

Diagnostic Utility of Thyrogen Whole Body Scanning Compared to Withdrawal Whole Body Scanning

The hypotheses to be tested were:

H_0: Thyrogen whole body scanning correctly diagnoses the presence or absence of cancer no more often than Withdrawal whole body scanning.

H_1: Thyrogen whole body scanning correctly diagnoses the presence or absence of cancer more often than Withdrawal whole body scanning.

Diagnostic Utility of Thyrogen Whole Body Scanning and Tg Testing Compared to Withdrawal Whole Body Scanning and Tg Testing

The hypotheses to be tested were:

H_0: Combining a Thyrogen WBS and a Tg test correctly diagnoses the presence or absence of cancer no more often than combining a Withdrawal WBS and a Tg test.

H_1: Combining a Thyrogen WBS and a Tg test correctly diagnoses the presence or absence of cancer more often than combining a Withdrawal WBS and a Tg test.

Quality of Life

The hypotheses to be tested were:

H_0: There is no difference in SF-36 QOL score after Thyrogen administration than during Withdrawal.

H_1: There is a difference in SF-36 QOL score after Thyrogen administration than during Withdrawal.

7.2.2 Sample Size

The following was presented by the sponsor for the sample size for the primary endpoint of comparing the proportion of discordant scan pairs within an arm. With a sample size of 100 evaluable patients per arm, an observed fraction of approximately 85% of the patients with Thyrogen scan class higher than or equal to the Withdrawal scan would give a 95% confidence interval of 76% to 91%, while an observed rate of 90% would give a 95% confidence interval of 82% to 95%.

For comparing the proportion of discordant scan pairs between the two arms of the study, 100 evaluable patients per arm would give 80% power in a one-tailed test at a critical level of 0.05 to detect a difference of 0.15 between proportions, e.g., 0.70 vs 0.85. An observed difference of 0.10 or larger would be statistically significant.

For the Billewicz scale assessments of hypothyroidism, 100 patients would give 80% power to test the hypothesis that patient hypothyroid symptoms are at least as pronounced after Thyrogen administration as during Withdrawal, against the alternative that patients after Thyrogen are on average at least 0.28 better on each symptom than during Withdrawal at a critical level of 0.05. A standard deviation of 0.958, as observed in Study TSH92-0601, was assumed. The sponsor recognized that the Wilcoxon Signed Rank test, which was used for the actual hypothesis testing, was not quite as efficient as the paired t-test.

One hundred and twenty patients per arm were expected to be enrolled to allow for up to 17% of the patients to be determined as not evaluable, as was found in Study TSH92-0601.

8.0 SPONSOR'S STATISTICAL ANALYSIS METHODS

8.1 Study TSH92-0601

8.1.1 Per Protocol Analyses

8.1.1.1 Primary Endpoint

All hypothesis testing was done at the 0.05 significance level.

Comparison of Scan Cancer Classification

If the sponsor determined that there was a sufficient number of discordant pairs, then the Sign Test was used to test the one-sided null hypothesis that the Thyrogen scan produces the higher staging at least as often as the Withdrawal scan. Otherwise, a 95% confidence was constructed for the proportion of scans in which the Thyrogen scan rating was higher than or clinically equivalent to the Withdrawal scan rating.

8.1.1.2 Secondary Endpoints

All secondary endpoint hypothesis testing was done at the 0.05 significance level.

Percent ^{131}I Uptake

The mean absolute and percent difference in ^{131}I uptake between the Thyrogen and Withdrawal scans was compared using the paired t-test.

Hypothyroid Symptoms

The number of patients reporting hypothyroid symptoms via Billewicz score at each time point measured was determined. Patients were classified as euthyroid, exhibiting hypothyroid symptoms, or hypothyroid according to this scale. The number of patients who exhibit a score of each of the POMS sub-scales which indicates an increase in hypothyroid symptoms was reported.

The mean or median change in symptoms from baseline to Thyrogen scan was compared to that of baseline to Withdrawal scan. The analysis of the paired data used the Signed Rank test for the Billewicz scale and the paired t-test for the POMS scale and its subscales.

8.1.1.3 Safety Data

No statistical methodology was presented for the display or analysis of adverse event or safety data.

8.1.2 *Post Hoc* Analyses

Analyses on the following variables were not defined in the protocol: thyroidal ^{131}I uptake, whole body ^{131}I retention, and serum ^{131}I levels, thyroglobulin response to Thyrogen, and thyroglobulin testing with Thyrogen, diagnostic operating characteristics for Thyrogen and Withdrawal scans (sensitivity, specificity, positive and negative predictive values, accuracy). These analyses will not be further discussed in this statistical review. Refer to the Medical Reviewer's report for clinical relevance.

8.2 Study TSH95-0101

Two patient groups were described for use in the efficacy analyses.

- The Intent-to-Treat (ITT) population was defined as all patients who were randomized and who had baseline and at least one efficacy evaluation after receiving at least one injection of Thyrogen. Primary, secondary, and tertiary efficacy analyses were performed on the ITT population.
- The Efficacy Evaluable (EFF) population was defined as all patients enrolled except those who: a) failed Inclusion/Exclusion criteria, b) failed TSH levels on first study day, c) failed TSH level at Withdrawal, d) failed criteria for radioiodine administration used for scan, e) failed to complete the study, and f) ineligible scan pairs for review. Only primary efficacy analyses were performed on the EFF population.

The population used in the safety analyses was defined as all patients who enrolled in the study and received at least one injection of Thyrogen.

8.2.1 Per Protocol Analyses

8.2.1.1 Demographic and Safety Data

Patient demographic data and baseline characteristics were presented by treatment arm and combined. Data were presented using frequencies, percentages and descriptive statistics. Treatment group comparisons were made using ANOVA adjusting for center effect. A test for homogeneity was done to verify comparability of treatment groups at randomization. All testing was performed at the 0.05 alpha level.

The significance of shifts from baseline in the laboratory data was evaluated within each treatment arm using McNemar's test for binary variables and the generalized McNemar's test for more than two categories. Descriptive statistics were presented for the laboratory data and vital sign data for each study day and for the change from baseline. These changes were assessed for each treatment arm using the paired t-test.

Adverse events data were presented by body system and preferred term based on the COSTART coding. The incidence adverse events was tabulated overall and by severity, relationship to study drug, and duration for each treatment arm.

8.2.1.2 Primary Endpoints

All primary endpoint hypothesis testing was done at the 0.05 significance level.

Comparison of Scan Cancer Classification Within Treatment Arm

Using data from the within patient comparison of the scans, the classification of cancer seen on the 48 hour Thyrogen scan was compared to the classification seen on the 48 hour Withdrawal scan. Two proportions were used for hypothesis testing: 1) The proportion of scans where the Thyrogen classification was higher than or equal to the Withdrawal classification; and 2) the proportion of scans where the Thyrogen classification was lower than the Withdrawal classification. These proportions were presented as both a point estimate and a 95% confidence interval. In addition, a two-tailed sign test was used to test whether the discordances significantly favor the Thyrogen or Withdrawal WBSs.

Comparison of Scan Cancer Classification Between Treatment Arms

The proportion of discordant scans where the Thyrogen WBS is classified higher than the Withdrawal WBS, θ, was calculated for each treatment arm. The two-tailed Fisher's Exact test was used to test whether Arm I was significantly different from Arm II.

The equivalence of scan cancer classification by treatment arm and between treatment arms analyses were also performed for the principal investigators' classification and McNemar's test along with a Kappa statistic and 95% confidence interval were used to evaluate the agreement between the blinded reader consensus results and the principal investigators results.

In addition, the blinded reader analyses were performed for the following 6 subgroups: patients with positive scan; recent thyroidectomy patients; post therapy (follow-up) patients; thyroid cancer class $=1$ versus class ≥ 2 patients; low risk versus high risk patients; and metastatic patients.

Hypothyroid Symptoms

Data from the Billewicz scale were used to determine if the patients after Thyrogen administration manifest significantly less pronounced symptoms of hypothyroidism than during Withdrawal. The median change in symptoms from baseline to Thyrogen scan was compared to that of baseline to Withdrawal scan using the Wilcoxon Signed Rank test within each treatment arm and the Mann Whitney test was used for the between arm comparison.

8.2.1.3 Secondary Endpoints

Diagnostic Utility Analyses

Various diagnostic utility analyses were to be performed for each treatment arm. The analyses were to present the prevalence, sensitivity, specificity, positive and negative predictive values, and accuracy of the various testing modalities compared to several reference standards. In addition, the receiver operator curve of each test was to be computed using the Tg level cut-offs mentioned below. Table 8.1 presents the per protocol diagnostic utility analyses.

Table 8.1
Per Protocol Diagnostic Utility Analysis (For the Detection of Metastatic Cancer Only)

Diagnostic Test Evaluated	Reference Standard Used	Reference Standard 1	Reference Standard 2
THST Tg level	1,2	Withdrawal post therapy scan class ≥ 2	Withdrawal post therapy scan class ≥ 2 OR a negative post therapy scan with Withdrawal Tg ≥ 10 ng/mL OR if no post therapy scan, a Withdrawal Tg ≥ 10 ng/mL and a decision by the physician to treat the patient (e.g. ^{131}I therapy, surgical dissection, external radiation, etc.).
Thyrogen Tg level	1,2		
Withdrawal Tg level	1		
Thyrogen Diagnostic WBS	2		
Withdrawal Diagnostic WBS	2		
Combination of a Thyrogen Diagnostic WBS and Tg level	1,2		
Combination of a Withdrawal Diagnostic WBS and Tg level	1		

Source: Figure 4 from Section 5.3: Statistical Methods and Analysis Plans, Volume 1.59, page 286.

All diagnostic utility analyses pertaining to Tg testing were limited to patients who were successfully ablated and Tg antibody negative. Each patient was classified into one of the following four groups for each reference standard and cut-off value: true positive, true negative, false positive, false negative.

Three separate Tg level cut-offs (≥2, 5, and 10 ng/mL) were to be used for the detection of well-differentiated thyroid cancer by a Tg test. All of these cut-offs have been reported in the literature to be indicators of cancer in patients who were successfully ablated. The Thyrogen stimulated Tg levels drawn on the day with the highest proportion of patients were above the cut-off for detecting thyroid cancer, i.e. ≥10 ng/mL, was to be assessed. A withdrawal Tg level under conditions of TSH stimulation of ≥10 ng/mL was considered indicative of thyroid cancer by the medical community. This was the cut-off that the sponsor had chosen.

All diagnostic utility analyses were conducted to include both high and low risk patients, as determined by the TNM (Tumor-Node-Metastasis) classification system, and to study these groups separately. 95% confidence intervals for sensitivity and specificity were calculated for each the diagnostic utility analyses listed below.

Diagnostic Utility of Thyrogen Tg Testing Compared to THST (Baseline) Tg Testing and Diagnostic Utility of Thyrogen Tg Testing Compared to Withdrawal Tg Testing
The one-sided sign test was used to test the hypothesis that Tg test with Thyrogen more often correctly diagnosed the presence or absence of cancer than a Tg test on THST and, separately, on Withdrawal. That is, Tg testing on Thyrogen has a higher sensitivity and/or specificity.

Diagnostic Utility of Thyrogen Whole Body Scanning Compared to Withdrawal Whole Body Scanning
The one-sided sign test was used to test the hypothesis that whole body scanning with Thyrogen more often correctly diagnoses the presence or absence of cancer than whole body scanning on Withdrawal. That is, whole body scanning on Thyrogen has a higher sensitivity and/or specificity.

Diagnostic Utility of Thyrogen Whole Body Scanning and Tg Testing Compared to Withdrawal Whole Body Scanning and Tg Testing
The one-sided sign test was used to test the hypothesis that combining a WBS and a Tg test with Thyrogen more often correctly diagnoses the presence or absence of cancer than combining a WBS and a Tg test on Withdrawal. That is, combining a WBS and a Tg test on Thyrogen has a higher sensitivity and/or specificity.

Quality of Life
The Wilcoxon Signed Rank test was used to test the change in SF-36 QOL score from baseline to Thyrogen scan compared to that of baseline to Withdrawal scan for each treatment arm. The Mann Whitney test was used for the between arm comparison. Figures were used to show the mean score of the SF-36 QOL instrument and the change from baseline by treatment arm.

8.2.2 *Post Hoc* Analyses
Study TSH95-0101, the sponsor's "confirmatory trial," had several analyses that were not defined in the protocol. All these analyses pertained to diagnostic utility characteristics.

8.2.2.1 Finding Comparable Thyrogen and Withdrawal Tg Levels
Using guidance from their investigators, the sponsor redesigned the diagnostic utility analysis using various Withdrawal reference standards. These reference standards consisted of a Withdrawal Tg level, ranging from 1 ng/mL to 10 ng/mL, or a positive Withdrawal diagnostic or post-therapy scan.

Receiver operator characteristic (ROC) curve analyses were then used to justify the "equivalence" of Thyrogen testing to Withdrawal testing for the diagnosis of remnants and cancer. The "equivalence" was based on finding "optimal" Thyrogen Tg values that resulted in an "optimal" balance between sensitivity and specificity that correlated well with the reference Withdrawal Tg value. The Withdrawal sensitivity and specificity values were arbitrarily chosen to be 100% for any one of the Withdrawal Tg values. The sponsor claimed that this was a more objective way to evaluate the "equivalence" of the two treatments.

The sponsor's analysis approach was as follows. The data was combined from both arms of the study. The rationale for combining the two arms was to have the largest sample size (n=141) for this analysis. The 72 hour Thyrogen Tg data, Withdrawal 48 hour Tg data, and 48 hour Withdrawal and Thyrogen scans were used. Thyrogen Tg cut-off levels of 1 to 10 ng/mL were evaluated for each of the Withdrawal reference standards. That is, an ROC curve with 10 points, one point for each of the 10 Thyrogen Tg cut-offs, for each of the 10 different Withdrawal Tg reference standards was calculated. The Withdrawal reference standards were as follow:

> Withdrawal Tg level ≥ Y (Y ranges from 1 to 10 ng/mL)**OR** a Withdrawal diagnostic or post-therapy scan ≥ 1.

A plot of all 10 ROC curves visually showed that the ROC curves for the Withdrawal reference standards with Tg levels of 2 and 10 ng/mL were separate from the other curves. All other curves were visually similar, that is, they were grouped together between the 2 and 10 ng/mL curves. The sponsor then arbitrarily chose the 5 ng/mL Withdrawal reference curve as the "medial" curve, representative of the group of curves. Next, for each of these three ROC curves (2, 5, and 10 ng/mL Withdrawal Tg reference), the "optimal" balance between sensitivity and specificity was chosen by